Machine learning

**Modul 6 Introducing Natural Language Processing (NLP)**

1 Module 6 : Introduction au traitement du langage naturel

**2 Aperçu du module**  
• Sections  
• Aperçu du traitement du langage naturel (NLP)  
• Services gérés de traitement du langage naturel  
• Conclusion du module  
• Démonstrations  
• Présentation d'Amazon Polly  
• Présentation d'Amazon Comprehend  
• Présentation d'Amazon Translate  
• Laboratoire  
• Laboratoire guidé : Créer un bot pour planifier des rendez-vous  
• Vérification des connaissances

Ce module comprend les sections suivantes :  
• **Introduction au traitement du langage naturel (NLP).** Cette introduction inclut une description des principaux défis rencontrés lors du travail avec le NLP et du processus global de développement des applications NLP.  
• Un examen de cinq services AWS que vous pouvez utiliser pour accélérer le développement des applications basées sur le NLP.

Ce module comprend également des démonstrations <enregistrées / dirigées par un formateur> qui vous montreront comment utiliser Amazon Comprehend, Amazon Polly et Amazon Translate en utilisant la console de gestion AWS.

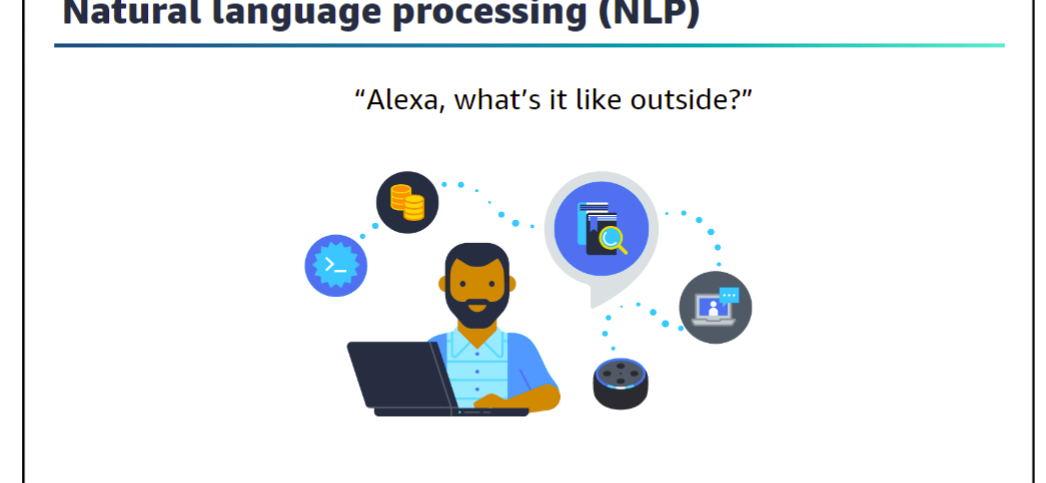
Le module comprend également un laboratoire pratique guidé où vous apprendrez à utiliser Amazon Lex pour développer un bot permettant de planifier des rendez-vous.

Enfin, vous devrez compléter une vérification des connaissances pour tester votre compréhension des concepts clés couverts dans ce module.

3 **Objectifs du module**  
• À la fin de ce module, vous devriez être capable de :  
• Décrire les cas d'utilisation du traitement du langage naturel (NLP) qui sont résolus en utilisant les services ML gérés d'Amazon  
• Décrire les services ML gérés d'Amazon disponibles pour le NLP  
• Utiliser les services ML gérés d'Amazon

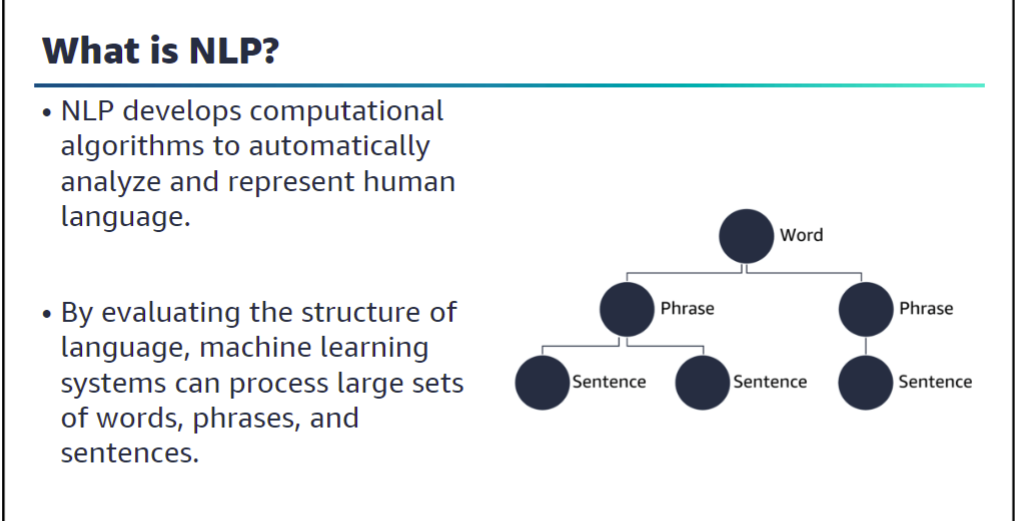
**Après avoir terminé ce module, vous devriez être capable de :  
• Décrire les cas d'utilisation du traitement du langage naturel** (NLP) qui sont résolus en utilisant les services ML gérés d'Amazon  
• Décrire les services ML gérés d'Amazon disponibles pour le NLP  
• Utiliser les services ML gérés d'Amazon

**Section 1 : Aperçu du traitement du langage naturel.**  
Dans cette section, vous examinerez la signification du traitement du langage naturel

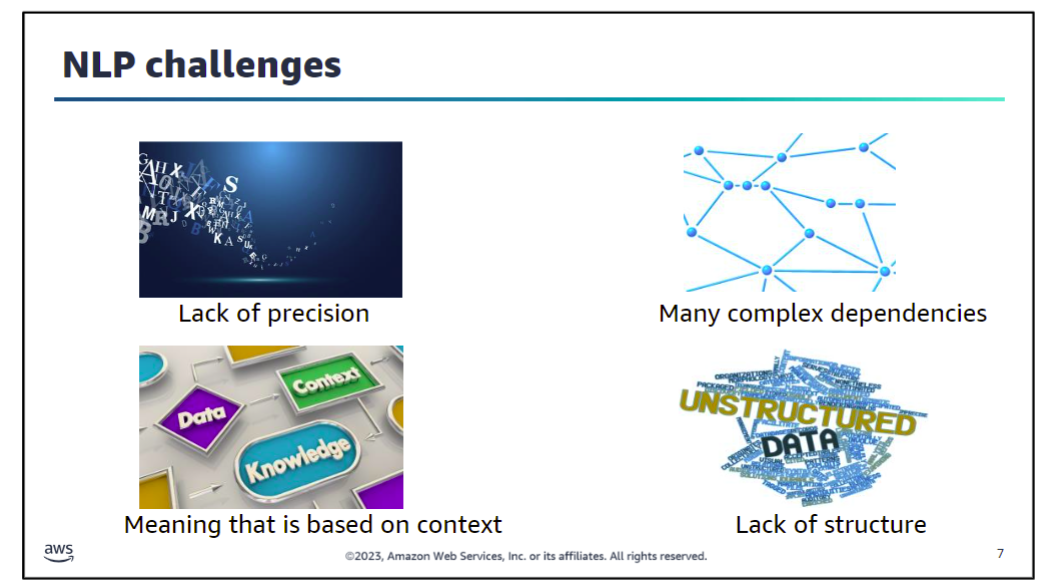


Avant de voir une explication du traitement du langage naturel (NLP), regardez un exemple de NLP avec Amazon Alexa.

1. Un appareil Amazon, tel qu'un Echo, enregistre vos paroles. L'enregistrement de votre discours est envoyé aux serveurs d'Amazon pour être analysé plus efficacement.
2. Amazon décompose votre phrase en sons individuels. Ensuite, il se connecte à une base de données qui contient les prononciations de divers mots pour trouver ceux qui correspondent le mieux à la combinaison des sons individuels.
3. Il identifie les mots importants pour comprendre les tâches à effectuer et exécuter les fonctions correspondantes. Par exemple, si Alexa remarque des mots comme "dehors" ou "température", elle ouvre l'application météo.
4. Les serveurs d'Amazon renvoient les informations à votre appareil, et Alexa peut parler.



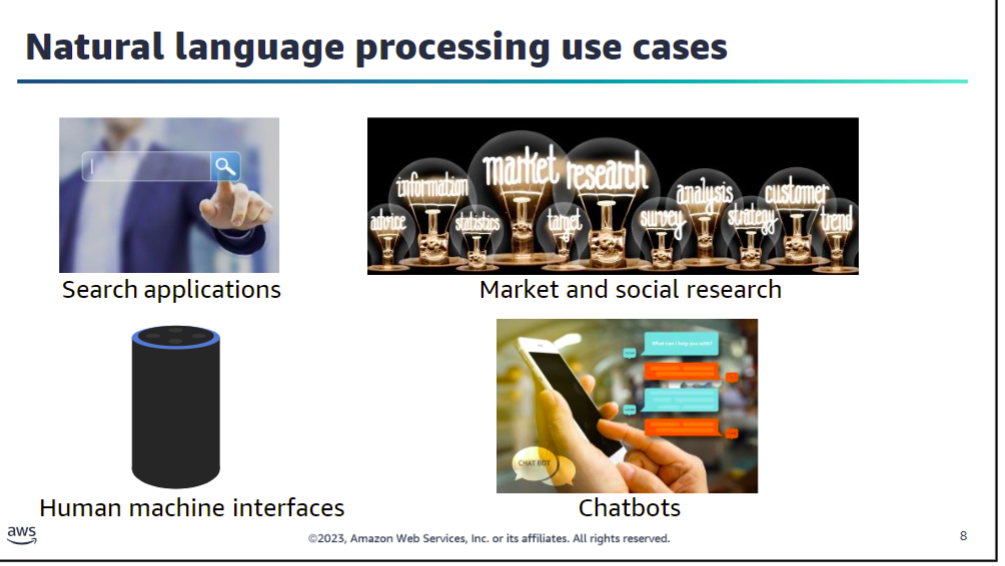
Le traitement du langage naturel (NLP) est un terme général qui désigne un ensemble de **problèmes commerciaux** ou **informatiques** que vous pouvez résoudre avec l'apprentissage automatique (ML). Les systèmes NLP existent avant le ML. Deux exemples sont la conversion de la parole en texte sur votre ancien téléphone portable et les lecteurs d'écran. De nombreux systèmes NLP utilisent maintenant une forme d'apprentissage automatique. Le NLP prend en compte la structure hiérarchique du langage. Les mots sont au niveau le plus bas de la hiérarchie. Un groupe de mots forme une phrase. Le niveau suivant est constitué des phrases, qui forment une phrase complète, et finalement, les phrases transmettent des idées. Les systèmes NLP sont confrontés à plusieurs défis majeurs, que vous apprendrez par la suite.



**Structure**  
Le langage n'est pas précis. Les mots peuvent avoir des significations différentes, qui dépendent des autres mots qui les entourent (contexte). Souvent, les mêmes mots ou expressions peuvent avoir plusieurs sens. Par exemple, prenez le terme *weather* (temps/météo). Vous pourriez être *under the weather*, ce qui signifie que vous êtes malade. Cependant, *there is wonderful weather outside* signifie que les conditions météorologiques à l'extérieur sont bonnes. L'expression *Oh, really?* peut exprimer la surprise, le désaccord ou de nombreuses autres significations, selon le contexte et l'intonation.

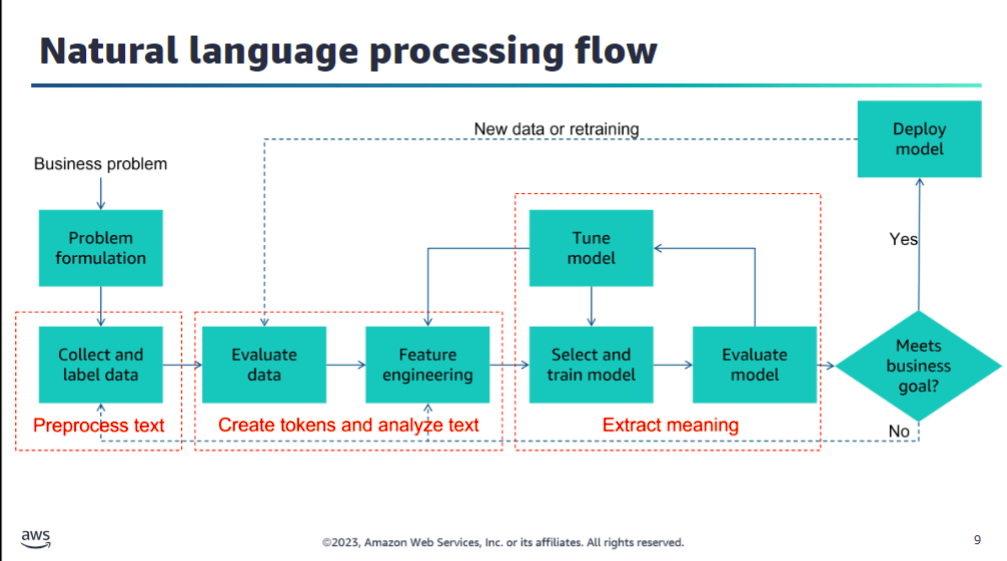
Certains **des principaux défis pour le traitement du langage naturel** (NLP) **incluent** :

* ***Découvrir la structure du texte*** – L'une des premières tâches de toute application NLP est de diviser le texte en unités significatives, telles que les mots, les phrases et les phrases complètes.
* **Étiquetage des données** – Après que le système a converti le texte en données, le défi suivant est d'appliquer des étiquettes qui représentent les différentes parties du discours. Chaque langue nécessite un schéma d'étiquetage différent pour correspondre à la grammaire de cette langue.
* **Représentation du contexte** – Puisque la signification des mots dépend du contexte, tout système NLP a besoin d'un moyen de représenter le contexte. C'est un grand défi en raison du grand nombre de contextes possibles. Convertir le contexte en une forme compréhensible par les ordinateurs est difficile.
* **Application de la grammaire** – Bien que la grammaire définisse une structure pour le langage, son application est presque infinie. Gérer la variation dans la manière dont les humains utilisent le langage est un défi majeur pour les systèmes NLP. C'est dans ce domaine que l'apprentissage automatique peut avoir un grand impact.

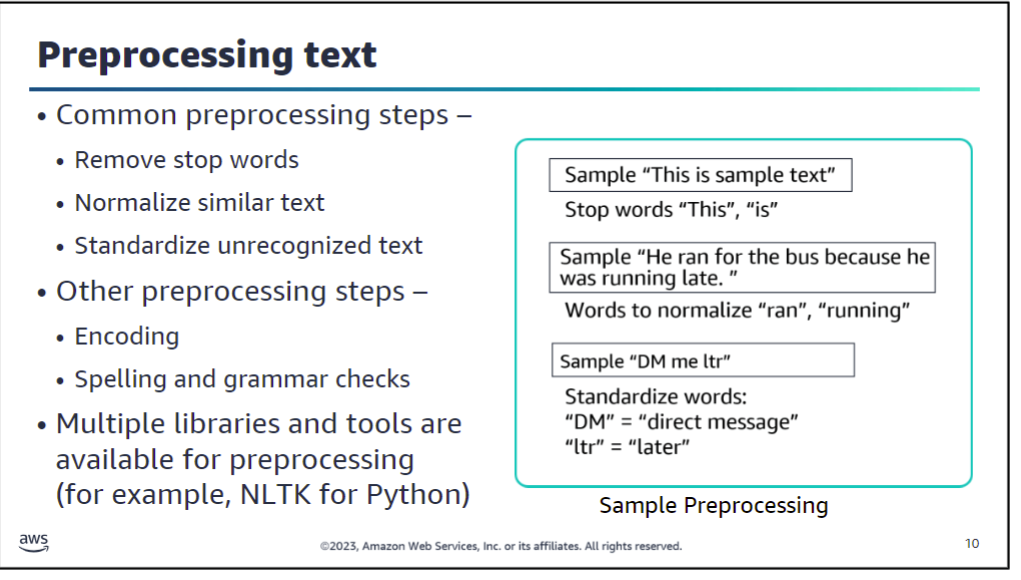


Vous pouvez appliquer le traitement du langage naturel (NLP) à un large éventail de problèmes. Certaines des applications les plus courantes incluent :

* *Applications de recherche (telles que Google et Bing)*
* *Interfaces homme-machine (telles qu'Alexa)*
* *Analyse des sentiments pour des campagnes marketing ou politiques*
* *Recherche sociale basée sur l'analyse des médias*
* *Chatbots qui imitent le discours humain dans des applications*
* *Flow- circulaire*



Vous pouvez appliquer le pipeline de développement d'apprentissage automatique (ML) que vous avez vu tout au long de ce cours lors du développement d'une solution NLP. La première tâche consiste à formuler un problème, puis à collecter et à étiqueter les données. Pour le NLP, la collecte de données consiste à diviser le texte en sous-ensembles significatifs et à étiqueter ces ensembles. L'ingénierie des caractéristiques est une grande partie des applications NLP. Ce processus devient plus compliqué lorsque vous travaillez avec du texte irrégulier ou non structuré. Par exemple, si vous développez une application pour classer des documents, vous devez être capable de distinguer les mots ayant des termes communs mais des significations différentes. L'étiquetage des données dans le domaine du NLP est parfois également appelé tagging (balisage). Dans le processus d'étiquetage, vous devez attribuer des chaînes de texte individuelles à différentes parties du discours. Vous pouvez utiliser des outils spécialisés pour aider à l'étiquetage dans le cadre du NLP.

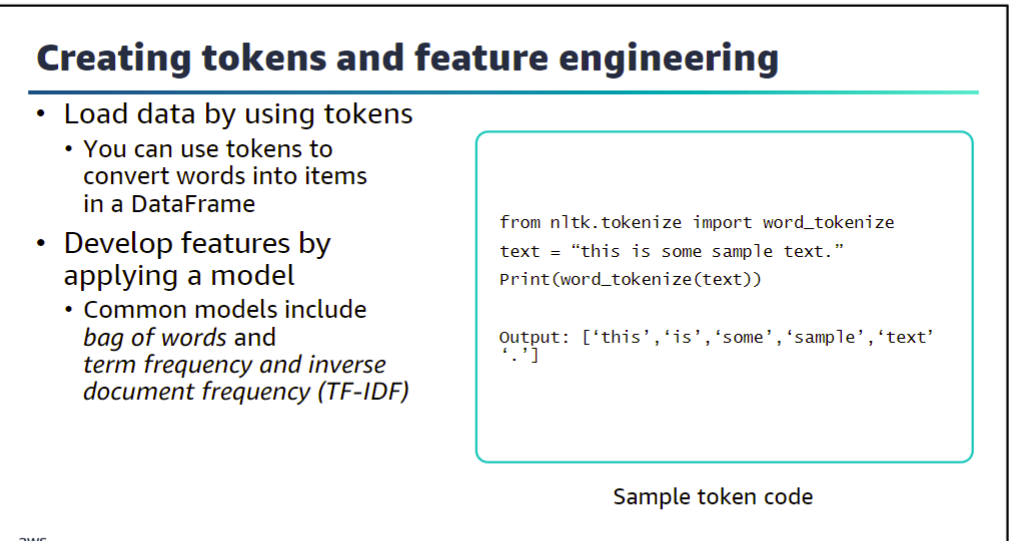


La première tâche pour une application NLP est de convertir le texte en données afin qu'il puisse être analysé. Vous convertissez le texte en supprimant les mots qui ne sont pas nécessaires pour l'analyse à partir du texte d'entrée. Dans l'exemple, les mots *This* et *is* sont supprimés pour ne laisser que l'expression *sample text*.

Après avoir supprimé ces mots vides, vous pouvez normaliser le texte en convertissant des mots similaires en une forme commune. Par exemple, les mots *run*, *runner*, *ran* et *running* sont tous des formes différentes du mot *run*. Vous pouvez normaliser toutes les occurrences de ces mots dans un bloc de texte en utilisant des processus de **racine** (stemming) et de **lemmatisation** (lemmatization).

Après avoir normalisé le texte, vous pouvez le standardiser en supprimant les mots qui ne figurent pas dans le dictionnaire que vous utilisez pour l'analyse. Des exemples incluent les acronymes, l'argot et les caractères spéciaux.

La bibliothèque Python *Natural Language Toolkit (NLTK)* fournit des fonctions que vous pouvez utiliser pour supprimer les mots vides, normaliser et standardiser le texte.

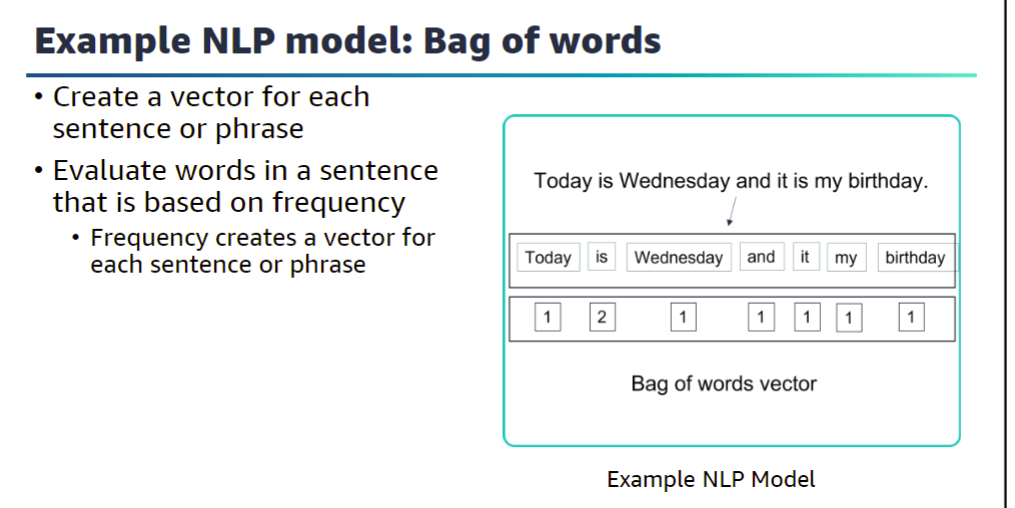


L'une des premières étapes pour créer un système NLP est de convertir le texte en une collection de données, telle qu'un DataFrame. Toutes les bibliothèques NLP fournissent des fonctions pour faciliter ce type de conversion. Cet exemple montre comment utiliser la fonction **word\_tokenize** fournie dans la bibliothèque NLTK.

Après avoir nettoyé votre texte et l'avoir chargé dans un DataFrame, vous pouvez appliquer l'un des modèles NLP pour créer des caractéristiques.

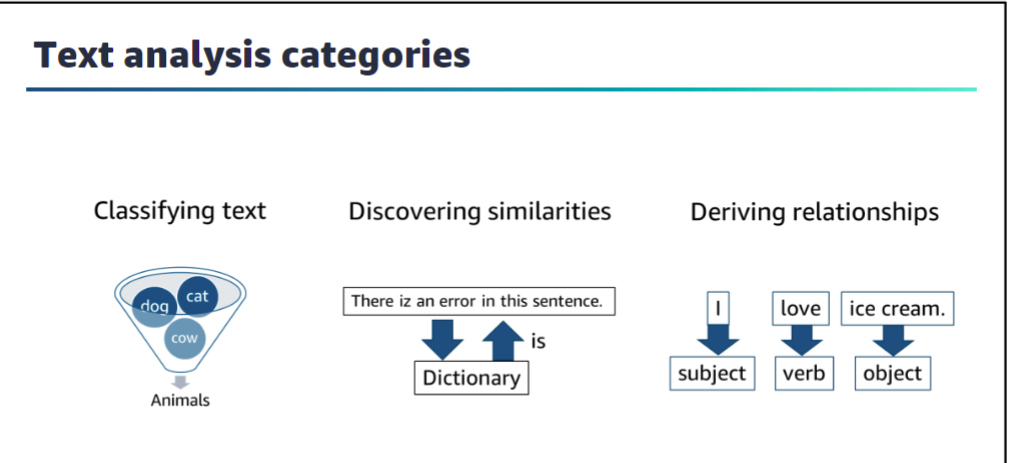
Les modèles courants incluent :

* **Sac de mots (Bag of words)** – Ce modèle simple capture la fréquence des mots dans un document. Pour chaque mot du document, une clé est créée, avec une valeur qui est le nombre d'occurrences dans ce document.
* **Fréquence des termes et fréquence inverse des documents (TF-IDF)** – La fréquence des termes est un décompte du nombre de fois qu'un mot apparaît dans un document. La fréquence inverse des documents est le nombre de fois qu'un mot apparaît dans un groupe de documents. Ces deux valeurs sont utilisées ensemble pour calculer un poids pour les mots. Les mots qui apparaissent fréquemment dans de nombreux documents ont un poids plus faible.



De nombreux modèles ont été établis dans le domaine du NLP. Cet exemple montre le modèle du sac de mots (bag of words). Le sac de mots est un modèle vectoriel. Les modèles vectoriels convertissent chaque phrase ou expression en un vecteur, qui est un objet mathématique enregistrant à la fois la directionnalité et la magnitude. Dans l'exemple, une phrase simple est convertie en un vecteur où chaque mot est enregistré en termes de fréquence. Le mot *is* a une valeur de 2 car il apparaît deux fois dans la phrase.

**Le sac de mots est souvent utilisé pour classer des documents dans différentes catégories. Il est également utilisé pour dériver des attributs qui alimentent des applications NLP, telles que l'analyse des sentiments**.

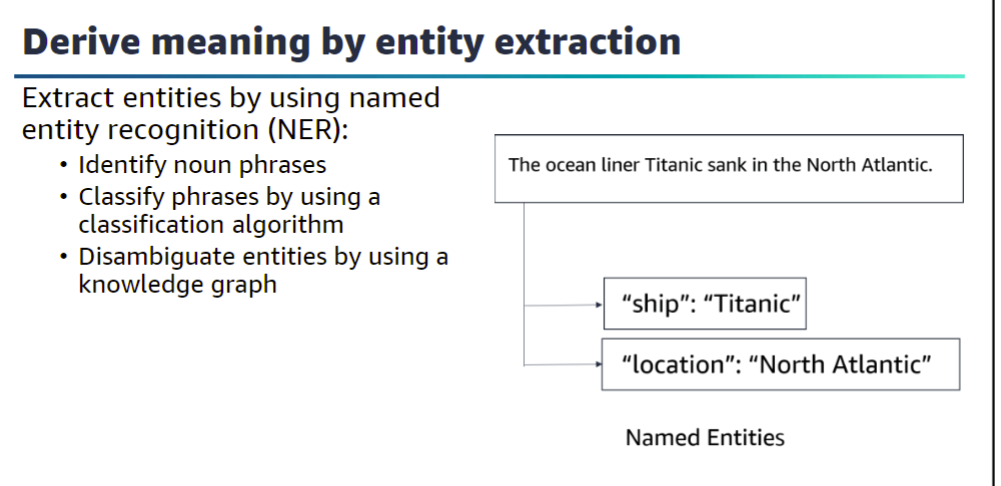


**L'analyse de texte se divise en trois grandes catégories :**

* **Classification du texte** – Cette catégorie d'analyse est similaire aux autres systèmes de classification que vous avez vus dans ce cours. Le texte fournit l'entrée à un processus qui extrait des caractéristiques, puis ces caractéristiques sont envoyées à travers un algorithme d'apprentissage automatique (ML). Cet algorithme interagit avec un modèle de classification pour inférer la classification. Vous pouvez **utiliser la bibliothèque Python NLTK pour** créer un système de classification.
* **Découverte des similitudes** – La correspondance de texte a de nombreuses applications. Par exemple, la correction automatique, la vérification orthographique et la vérification grammaticale sont toutes basées sur la correspondance de texte. L'algorithme de distance d'édition (également connu sous le nom de distance de Levenshtein) est fréquemment utilisé.
* **Dérivation des relations** – Vous pouvez dériver des relations entre différents mots ou phrases dans le texte en utilisant un processus appelé résolution des coréférences. Plusieurs systèmes NLP fournissent des bibliothèques Python pour dériver les relations.



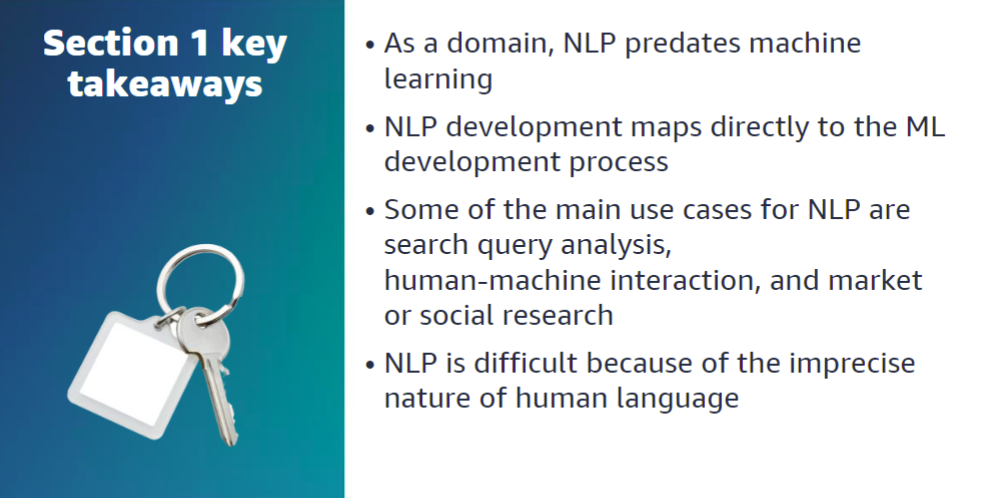
**Étiquetage**  
**L'un des plus grands défis pour le traitement du langage naturel (NLP) est de décrire le contexte du texte**. Prenons l'exemple où un utilisateur recherche le terme *tablette*. Comme le mot *tablette* a au moins deux significations distinctes, le moteur de recherche doit comprendre quel sens l'utilisateur a en tête. Le terme peut être précisé davantage (par exemple, en ajoutant un autre terme comme *médicament* ou *informatique*). S'il n'est pas précisé par d'autres termes, la plupart des moteurs de recherche se basent sur le contexte le plus couramment utilisé.



**Graphique**  
Le processus d'extraction des entités est appelé **reconnaissance des entités nommées (NER)**. Un modèle NER comporte les composants suivants :

* Identifier les syntagmes nominaux en utilisant des tableaux de dépendance et l'étiquetage des parties du discours.
* Classifier les syntagmes en utilisant un algorithme de classification, tel que Word2Vec.
* Désambiguïser les entités en utilisant un **graphique de connaissances**.

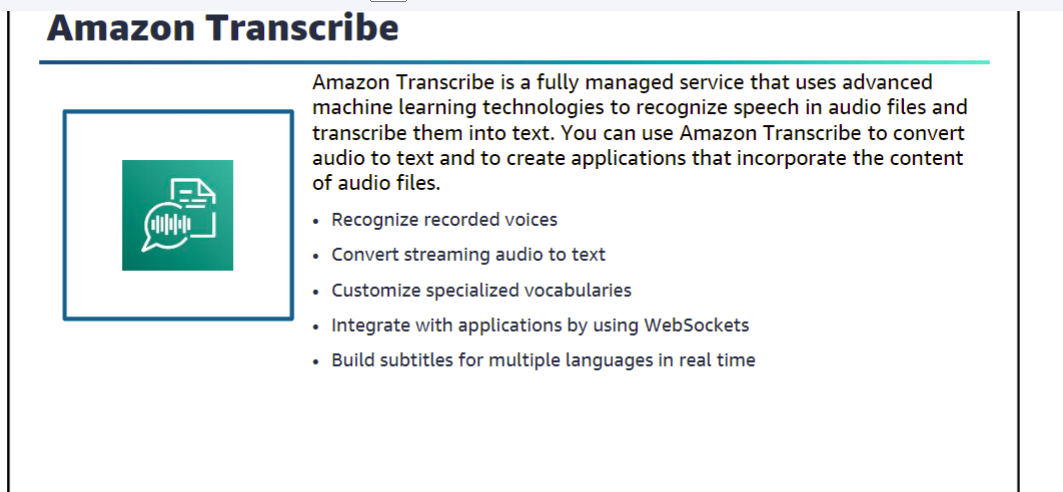
Cet exemple montre comment utiliser NER pour extraire les entités *Titanic* et *Atlantique Nord* à partir du texte. Une fois que les entités nommées sont extraites, vous pouvez utiliser un graphique de connaissances pour extraire le sens. Un graphique de connaissances combine l'expertise du domaine avec l'apprentissage automatique pour en tirer du sens. Le moteur de recommandations d'Amazon est un exemple de graphique de connaissances.



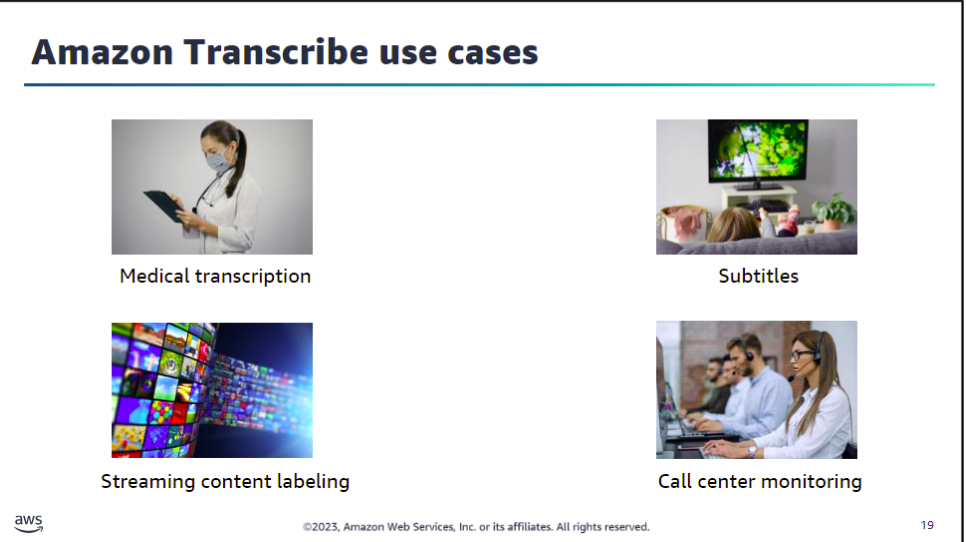
Quelques points clés à retenir de cette section du module incluent :

* En tant que domaine, le NLP (traitement du langage naturel) est antérieur à l'apprentissage automatique (ML).
* Le développement du NLP correspond directement au processus de développement de l'apprentissage automatique.
* Certains des principaux cas d'utilisation du NLP sont l'analyse des requêtes de recherche, l'interaction homme-machine, et la recherche marketing ou sociale.
* Le NLP est difficile en raison de la nature imprécise du langage humain.

**Section 2 : Services gérés de traitement du langage naturel**.  
Dans cette section, vous passerez en revue cinq services gérés d'apprentissage automatique (ML) d'Amazon que vous pouvez utiliser pour divers cas d'utilisation du NLP. Ces services simplifient le processus de création d'une application NLP.

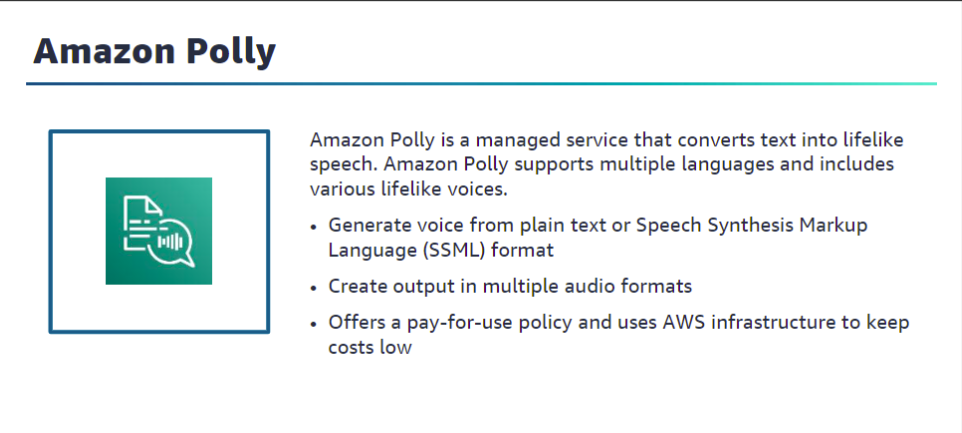


**Amazon Transcribe** est le premier service géré d'apprentissage automatique que vous allez découvrir. Vous pouvez utiliser Amazon Transcribe pour reconnaître la parole dans des fichiers audio et produire une transcription. Amazon Transcribe peut reconnaître des voix spécifiques dans un fichier audio, et vous pouvez créer un vocabulaire personnalisé pour des termes spécialisés dans un domaine particulier. Vous pouvez également ajouter le service de transcription à vos applications en intégrant des WebSockets. Les WebSockets offrent une interface orientée vers Internet qui permet une communication bidirectionnelle entre une application et Amazon Transcribe.



Voici quelques cas d'utilisation courants d'Amazon Transcribe :

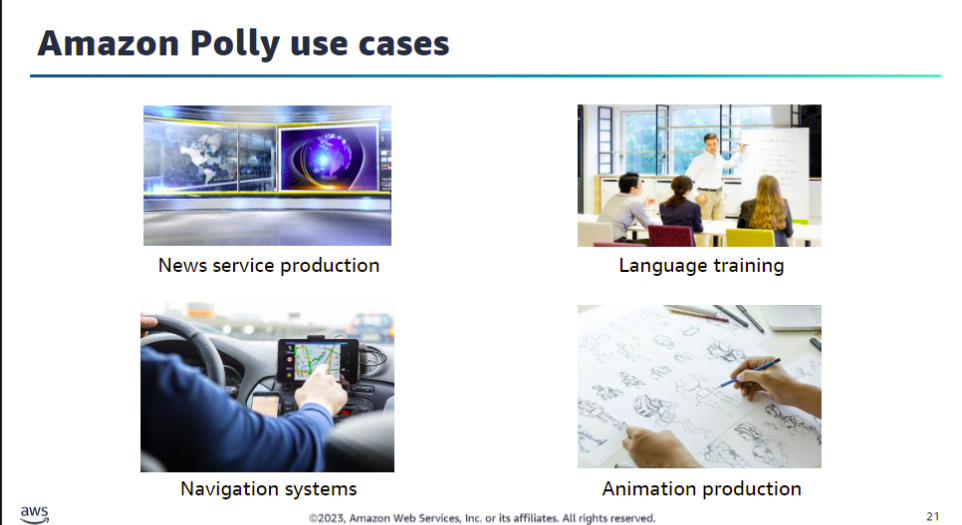
* **Transcription médicale** : Les professionnels de la santé peuvent enregistrer leurs notes, et Amazon Transcribe peut capturer ces notes vocales sous forme de texte.
* **Sous-titres pour vidéos** : Les organisations de production vidéo peuvent générer automatiquement des sous-titres à partir de vidéos. Cela peut également être fait en temps réel pour un flux en direct afin d'ajouter des sous-titres codés (CC).
* **Étiquetage de contenu en streaming** : Les entreprises de médias peuvent capturer et étiqueter du contenu, puis envoyer ce contenu à Amazon Comprehend pour une analyse plus approfondie.
* **Surveillance des centres d'appels clients** : Les entreprises peuvent enregistrer les appels de service client ou de vente, puis analyser les résultats pour des opportunités de formation ou des stratégies.



**Amazon Polly peut** *convertir du texte en une voix réaliste*. Vous pouvez saisir des fichiers en texte brut ou un fichier au format SSML (Speech Synthesis Markup Language). SSML est un langage de balisage que vous pouvez utiliser pour fournir des instructions spécifiques sur la manière dont la parole doit être restituée. Par exemple, vous pourriez vouloir **introduire une pause dans le flux de parole. Vous pouvez ajouter une balise SSML pour demander à Amazon Polly de faire une pause entre deux mots.**

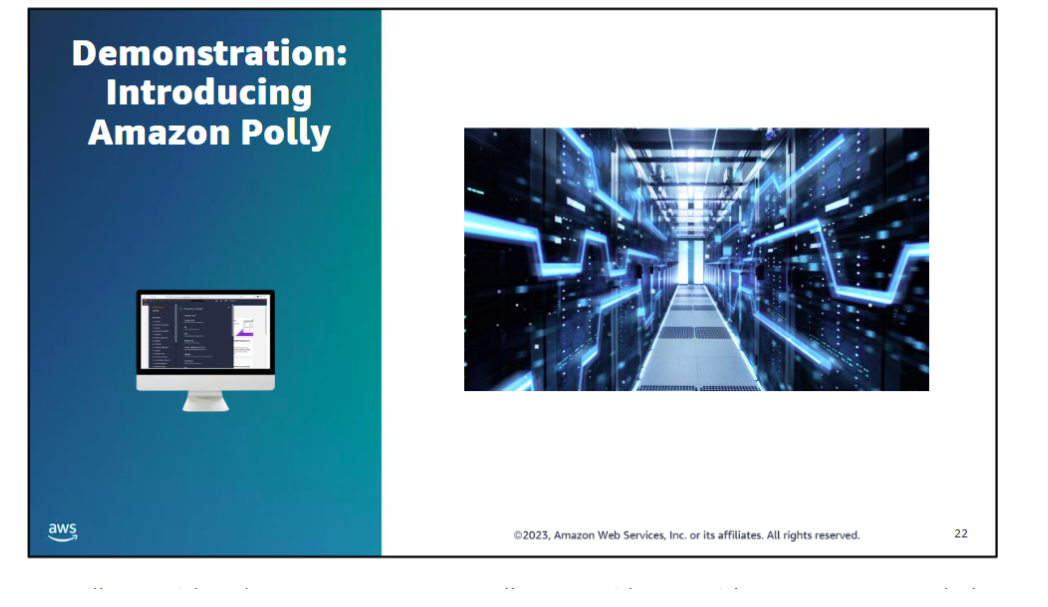
Vous pouvez également produire la parole générée par Amazon Polly dans les formats de flux audio MP3, Vorbis, et PCM (modulation par impulsion codée). Amazon Polly a diverses applications. Les cas d'utilisation courants incluent les applications mobiles (telles que les lecteurs d'actualités), les jeux, les plateformes d'apprentissage en ligne, et les applications d'accessibilité pour les personnes malvoyantes.

Amazon Polly est éligible à l'utilisation avec des charges de travail réglementées pour la loi américaine Health Insurance Portability and Accountability Act de 1996 (HIPAA) et la norme de sécurité des données de l'industrie des cartes de paiement (PCI DSS).

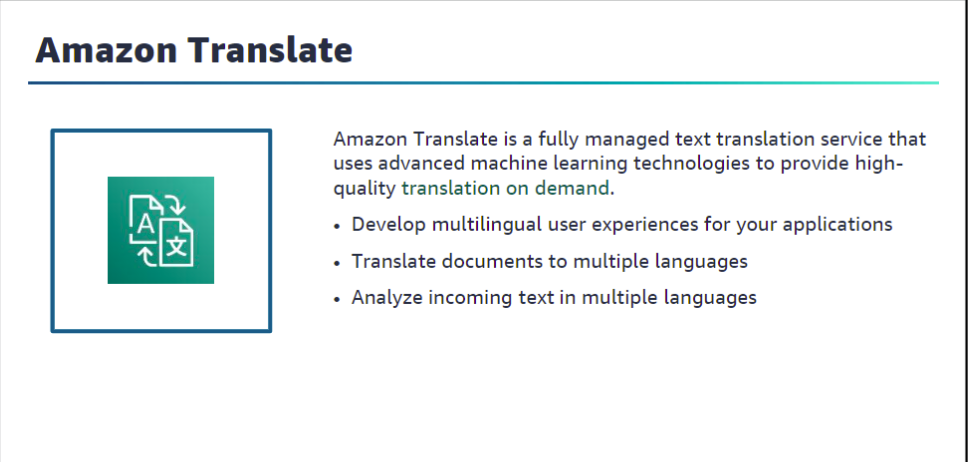


Voici quelques cas d'utilisation courants d'Amazon Polly :

* **Production de services d'actualités** : Les grandes entreprises de presse utilisent Amazon Polly pour générer du contenu vocal directement à partir de leurs articles écrits.
* **Systèmes de formation linguistique** : Les entreprises de formation linguistique utilisent Amazon Polly pour créer des systèmes d'apprentissage d'une nouvelle langue.
* **Systèmes de navigation** : Amazon Polly est intégré dans les interfaces de programmation d'applications de cartographie (API), permettant ainsi aux développeurs d'ajouter la voix à leurs applications géolocalisées.
* **Production d'animation** : Les animateurs utilisent Amazon Polly pour ajouter des voix à leurs personnages.



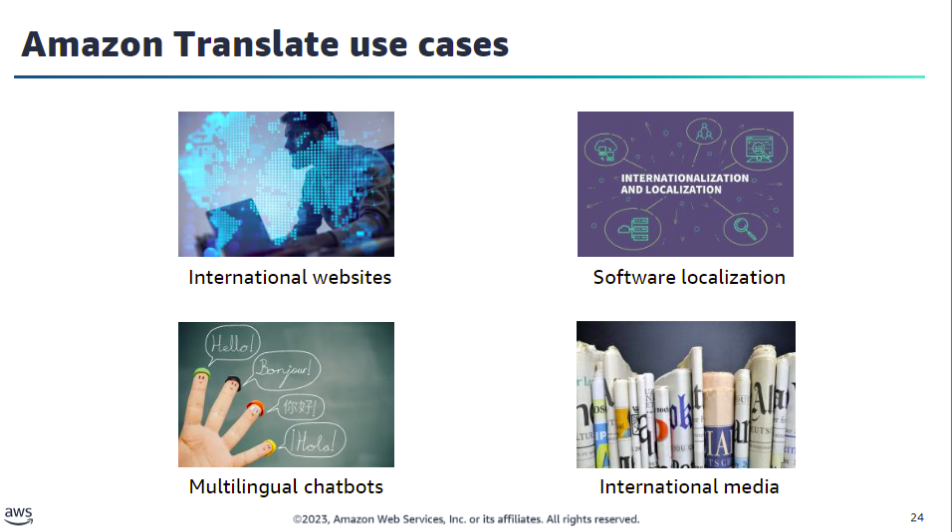
Votre instructeur va maintenant soit faire une démonstration d'Amazon Polly, soit vous donner accès à une démonstration enregistrée.



Avec **Amazon Translate**, vous pouvez créer des expériences multilingues dans vos applications. Vous pouvez concevoir des systèmes pour lire des documents dans une langue, puis les rendre ou les stocker dans une autre langue. Vous pouvez également utiliser Amazon Translate dans le cadre d'un système d'analyse de documents.

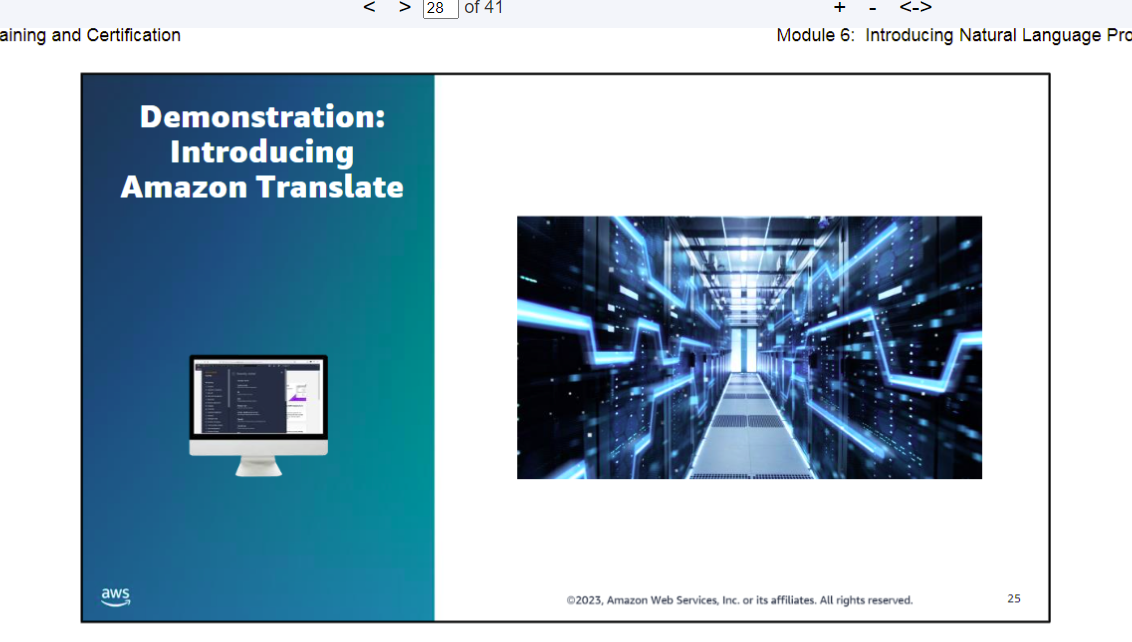
**Amazon Translate est entièrement intégré avec d'autres services de machine learning d'Amazon, tels qu'Amazon Comprehend, Amazon Transcribe et Amazon Polly. Grâce à cette intégration, vous pouvez :**

* Extraire des entités nommées, des sentiments et des phrases clés en intégrant avec Amazon Comprehend.
* Créer des sous-titres multilingues avec Amazon Transcribe.
* Faire parler du contenu traduit avec Amazon Polly.

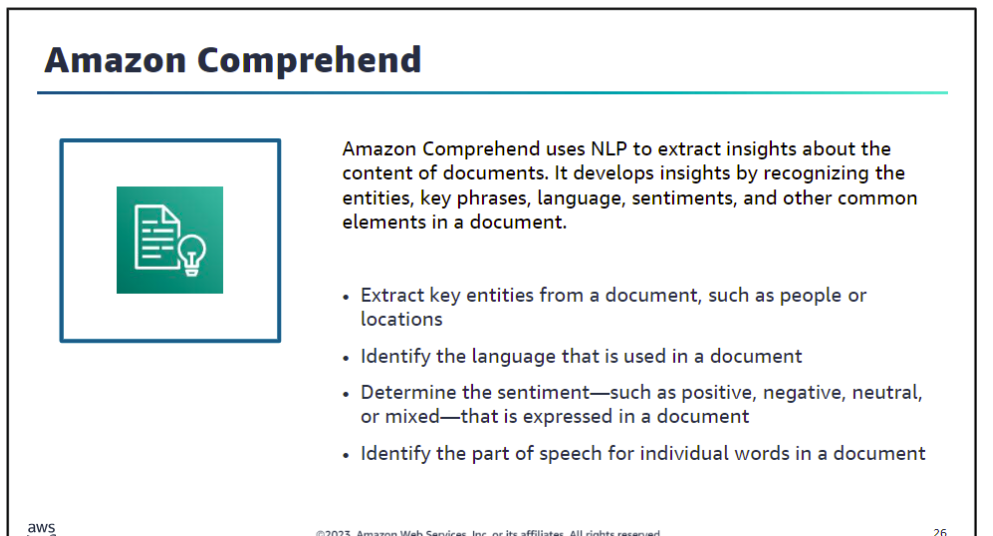


Voici quelques cas d'utilisation courants d'Amazon Translate :

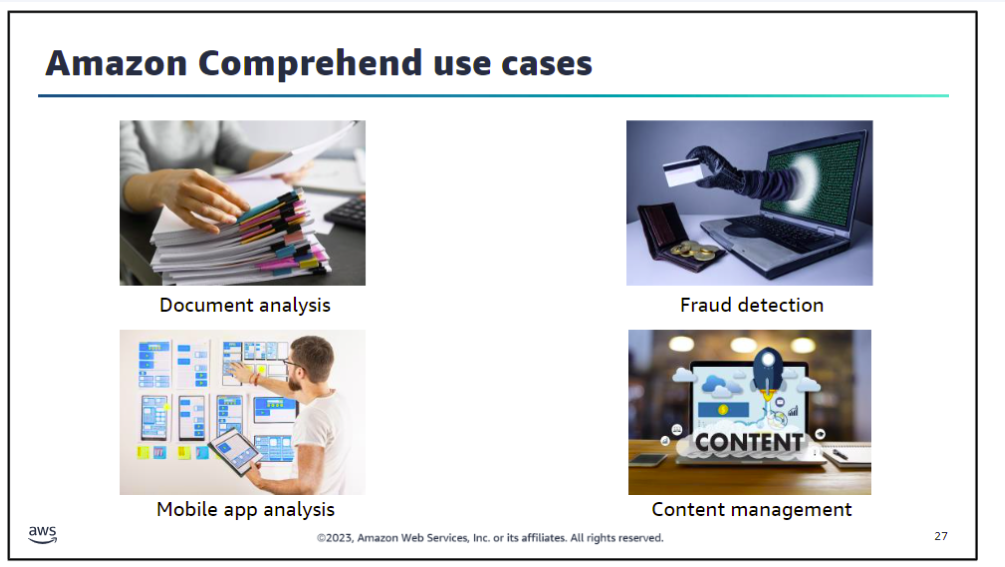
* **Sites web internationaux** : Vous pouvez utiliser Amazon Translate pour globaliser rapidement vos sites web.
* **Localisation de logiciels** : La localisation est un coût majeur pour tout logiciel destiné à un public mondial. Amazon Translate peut réduire le temps de développement de logiciels et diminuer considérablement les coûts de localisation.
* **Chatbots multilingues** : Les chatbots sont utilisés pour créer une interface plus humaine avec les applications. Avec Amazon Translate, vous pouvez créer un chatbot qui parle plusieurs langues.
* **Gestion internationale des médias** : Les entreprises qui gèrent des médias pour un public mondial utilisent Amazon Translate pour réduire leurs coûts de localisation.



Your instructor will now either demonstrate Amazon Translate or provide you with access to a recorded demonstration.

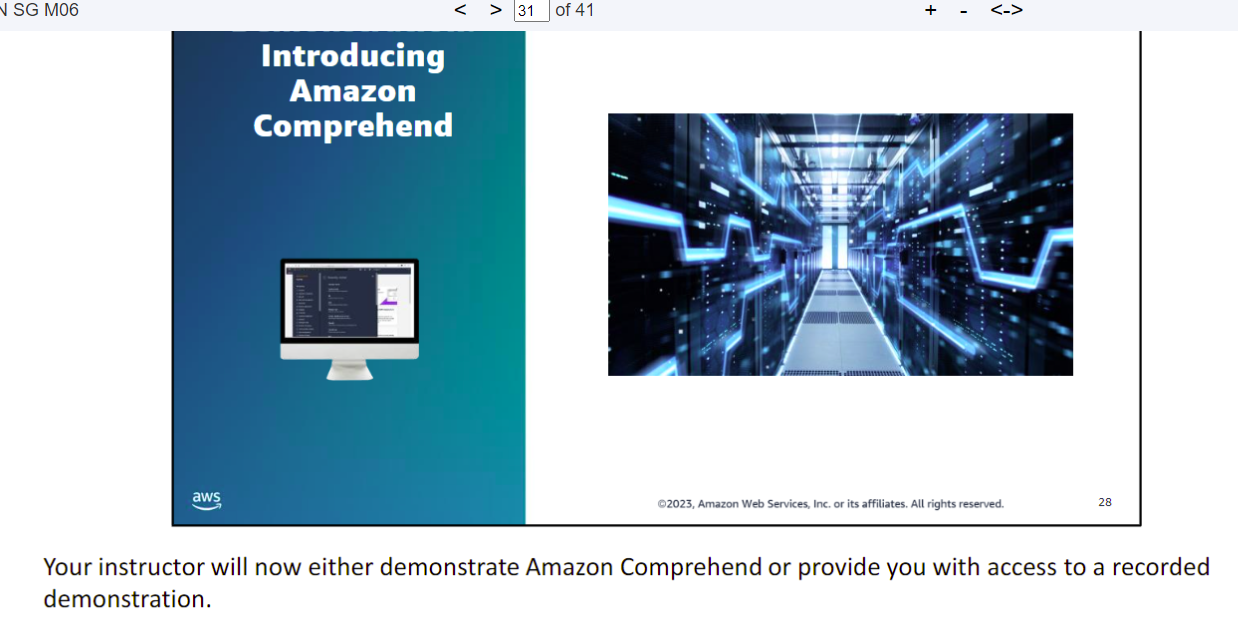


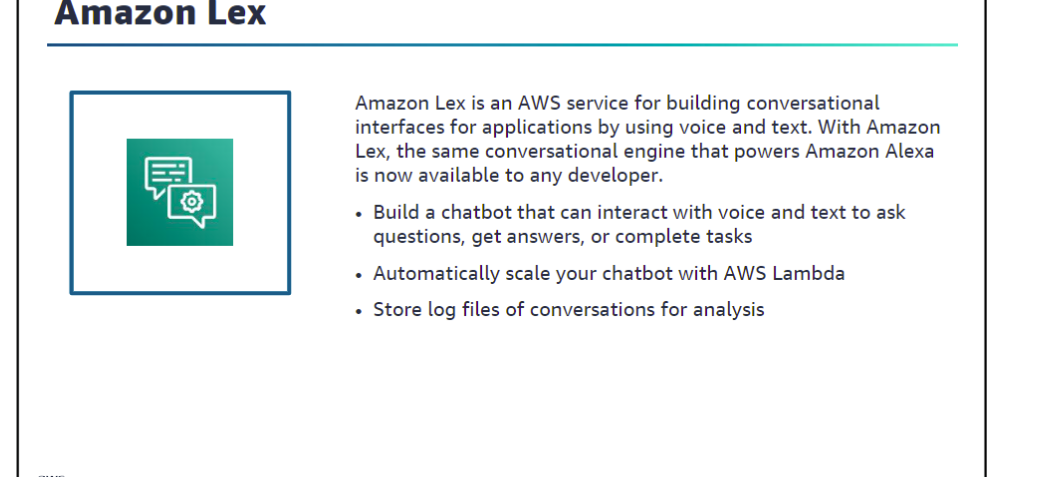
**Amazon Comprehend** met en œuvre de nombreuses techniques de traitement du langage naturel (NLP) que vous avez étudiées plus tôt dans ce module. Vous pouvez extraire des entités clés, effectuer une analyse de sentiments et étiqueter des mots avec des catégories grammaticales.



Voici quelques cas d'utilisation courants d'Amazon Comprehend :

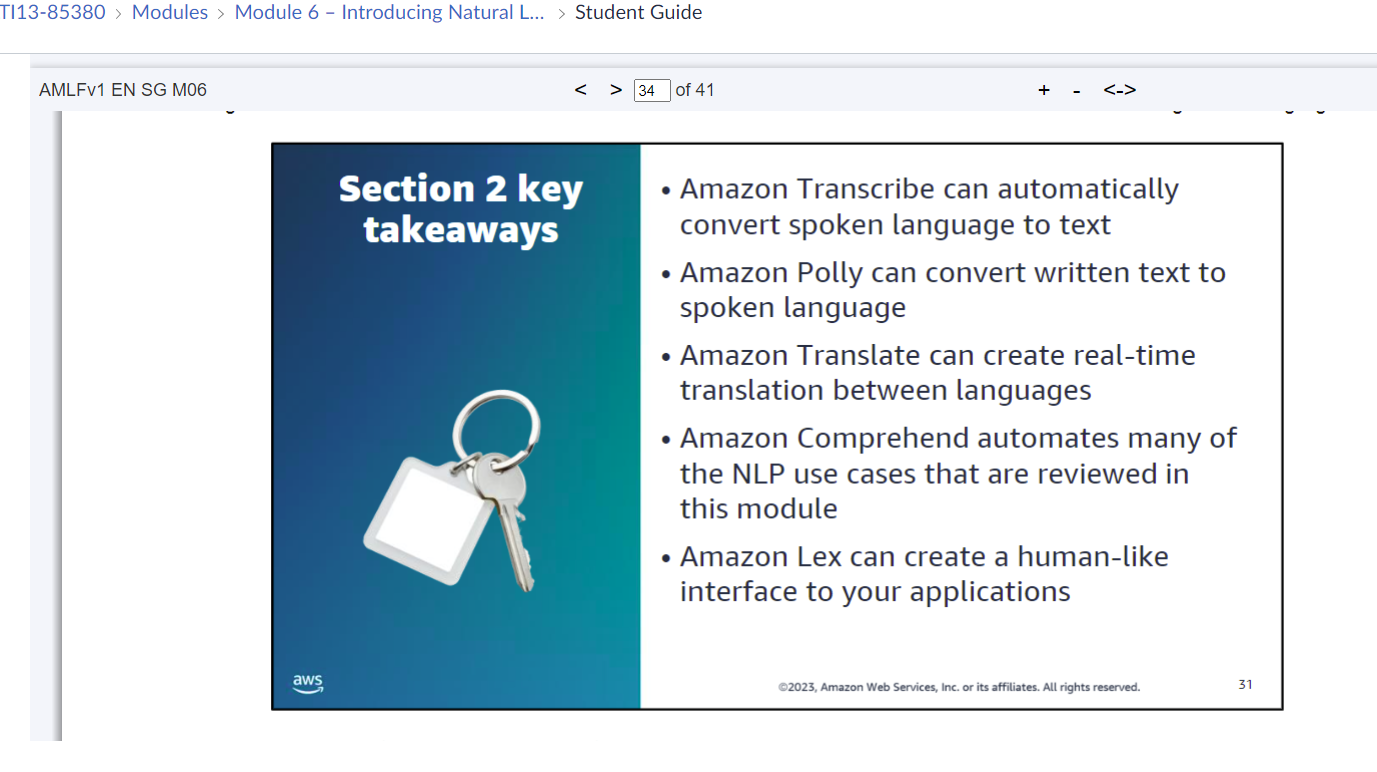
* **Analyse de documents juridiques et médicaux** : Les organisations juridiques, d'assurance et médicales ont utilisé Amazon Comprehend pour exécuter de nombreuses fonctions de traitement du langage naturel (NLP) abordées dans ce module.
* **Détection de fraude financière** : Les banques, institutions financières et autres organismes ont utilisé Amazon Comprehend pour examiner de grands ensembles de données de transactions financières afin de détecter la fraude et de repérer des schémas de transactions illégales.
* **Analyse d'applications mobiles à grande échelle** : Les développeurs d'applications mobiles peuvent utiliser Amazon Comprehend pour identifier des schémas dans l'utilisation de leurs applications afin de concevoir des améliorations.
* **Gestion de contenu** : Les entreprises de médias et autres sociétés de contenu peuvent utiliser Amazon Comprehend pour étiqueter du contenu à des fins d'analyse et de gestion.



Avec **Amazon Lex,** vous pouvez ajouter une interface en langage humain à vos applications. Amazon Lex vous permet d'utiliser le même moteur conversationnel qui alimente Amazon Alexa. Vous pouvez augmenter automatiquement la capacité en créant des fonctions AWS Lambda pour évoluer à la demande. De plus, vous pouvez stocker des fichiers journaux des conversations pour une analyse ultérieure.



*Certaines des cas d'utilisation les plus courants d'Amazon Lex incluent* :  
• **Création d'interfaces frontend pour la gestion des stocks et les ventes** – Les interfaces vocales deviennent de plus en plus courantes. *Les entreprises utilisent Amazon Lex pour ajouter des chatbots à leurs applications de gestion des stocks et de vente*.  
• **Développement d'assistants interactifs** – En combinant Amazon Lex avec d'autres services de machine learning, les clients créent des assistants plus sophistiqués pour de nombreux secteurs différents.  
• **Création d'interfaces de service client** – Les applications vocales similaires à celles des humains deviennent rapidement la norme pour les applications de service client. Amazon Lex peut réduire le temps de réponse et améliorer la qualité de ces chatbots.  
• **Interrogation de bases de données avec un langage semblable à celui des humains** – Amazon Lex est combiné avec d'autres services de base de données AWS pour créer des applications d'analyse de données sophistiquées avec une interface en langage humain.



Quelques points clés à retenir de cette section du module incluent :  
• **Amazon Transcribe** peut automatiquement convertir le langage parlé en texte.  
• **Amazon Polly** peut convertir le texte écrit en langage parlé.  
• **Amazon Translate** peut créer des traductions en temps réel entre les langues.  
• **Amazon Comprehend** automatise de nombreux cas d'utilisation du traitement du langage naturel (NLP) examinés dans ce module.  
• **Amazon Lex** peut créer une interface semblable à celle des humains pour vos applications.

We’ll get started by – nous allons commencer par

