

## Capítulo 5. Aplicación y adopción de estándares de anotación

Ahora que ha creado la especificación para su objetivo de anotación, está casi listo para comenzar a anotar su corpus. Sin embargo, antes de comenzar a anotar, debe considerar qué forma tomarán sus datos anotados, es decir, usted sabe *lo que* quiere que hagan sus anotadores, pero debe decidir *cómo* quiere que lo hagan. En este capítulo, examinaremos los diferentes formatos que puede tomar la anotación y discutiremos los pros y los contras de cada uno respondiendo las siguientes preguntas:

- ¿Cómo se ve la anotación?
- ¿Los diferentes tipos de tareas se representan de manera diferente?  
¿Si es así, cómo?
- ¿Cómo puede asegurarse de que su anotación pueda ser utilizada por otras personas y en conjunto con otras tareas?
- ¿Qué consideraciones se toman para decidir sobre un entorno de anotación y un formato de datos, tanto para los anotadores como para el aprendizaje automático?

Antes de entrar en los detalles de cómo aplicar su especificación a su corpus, debe comprender cómo se ve realmente la anotación cuando se ha aplicado a un documento o texto. Así que ahora veamos los ejemplos de especificaciones del [Capítulo 4](#) y veamos cómo se pueden aplicar a un corpus real.

Hay muchas maneras diferentes de representar la información sobre un corpus. Los ejemplos que le mostramos no serán exhaustivos, pero le brindarán una descripción general de algunos de los diferentes formatos que pueden tomar los datos anotados.

---

### NOTA

*Mantenga sus datos accesibles.* Su proyecto de anotación será mucho más fácil de administrar si elige un formato para sus datos que sea fácil de modificar y acceder. El uso de sistemas de bases de datos intrincados o esquemas XML complicados para definir sus datos está bien si está acostumbrado a ellos, pero si no lo está, será mejor que mantenga las cosas simples.

---

Las tareas de anotación van desde el simple etiquetado de documentos hasta el etiquetado de extensiones de texto y la vinculación de etiquetas. A medida que las especificaciones y las tareas aumentan en complejidad,

es necesario incluir más información en la anotación. En las siguientes secciones, analizaremos las formas más comunes en que estas tareas se representan en los datos y las ventajas y desventajas de cada estilo.

## Anotación de metadatos: clasificación de documentos

En el [Capítulo 4](#) analizamos un ejemplo de una tarea de clasificación de documentos, la de etiquetar los géneros de una película en función de un resumen o una reseña mediante el uso de etiquetas de categoría no exclusivas (una película puede ser tanto una comedia como un western, por ejemplo). Sin embargo, antes de llegar a varias etiquetas de categoría para un documento, veamos un ejemplo un poco más simple: etiquetar las reseñas de películas como positivas, negativas o neutrales con respecto a la película que están reseñando. Este es un ejercicio de categorización más simple porque las etiquetas no se superpondrán; cada documento tendrá una sola clasificación.

### Etiquetas únicas: Reseñas de películas

Supongamos que tiene un corpus de 100 reseñas de películas, con cantidades aproximadamente iguales de documentos positivos, negativos y neutrales. Al leer cada documento, usted (o sus anotadores) pueden determinar en qué categoría debe etiquetarse cada documento, pero ¿cómo va a representar esa información? Aquí hay algunas sugerencias de lo que puede hacer:

- Tener un archivo de texto u otro formato de archivo simple (p. ej., separados por comas) que contenga una lista de nombres de archivo y su etiqueta asociada.
- Cree un archivo de base de datos y haga que sus anotadores ingresen comandos SQL para agregar archivos a la tabla adecuada.
- Cree una carpeta para cada etiqueta en su computadora y, a medida que se clasifique cada revisión, mueva el archivo a la carpeta correspondiente.
- Pida a los anotadores que cambien el nombre del archivo para agregar la palabra *positivo* , *negativo* o *neutral* , como en *review0056-positive.txt* .
- Agregue la clasificación dentro del archivo que contiene la revisión.

Tenga en cuenta que estas opciones van desde representaciones completamente externas de los datos de anotación, donde la información se almacena en archivos completamente diferentes, hasta completamente internas, donde la información se mantiene dentro del mismo documento.

También cubren el término medio, donde la información se guarda en el sistema de archivos, cerca del corpus pero no como parte completa de él.

Entonces, ¿cuál de estos sistemas es el mejor? En [el Capítulo 4](#) , discutimos la importancia del estándar LAF y explicamos por qué la anotación independiente es preferible a realizar cambios en el texto real del corpus. Entonces, la última opción en la lista no es preferible.

Pero, ¿cómo eliges entre las otras cuatro opciones? Todos registran la información de la anotación mientras conservan el formato de los datos; es uno realmente mejor que el otro? En cuanto a la aplicación del modelo a los datos, diríamos que no, que no existe una diferencia real entre ninguna de las opciones restantes. Cada representación podría convertirse en cualquiera de las otras sin pérdida de datos o demasiado esfuerzo (suponiendo que usted o alguien que conozca pueda hacer alguna programación básica, o esté dispuesto a reformatear a mano).

Entonces, la decisión real aquí se basará en otros factores, como lo que será más fácil para sus anotadores y lo que dará como resultado las anotaciones más precisas. Pedirle a sus anotadores que aprendan los comandos SQL para crear tablas podría ser la mejor opción desde su perspectiva, pero a menos que sus anotadores ya estén familiarizados con ese lenguaje y la interfaz que lo acompaña, es probable que el uso de dicho sistema retrase en gran medida el proceso de anotación. y posiblemente resulte en anotaciones inexactas o incluso en la pérdida de datos si alguien logra eliminar su base de datos.

---

#### ADVERTENCIA

*¡Cuidado con las fuentes de error!* Las tareas de anotación suelen ser laboriosas y requieren atención a los detalles, por lo que cualquier fuente de confusión o error probablemente surgirá al menos unas cuantas veces.

---

Hacer que sus anotadores escriban información también puede ser problemático, incluso con una tarea de etiquetado simple como esta. Considere dar a sus anotadores una carpeta de archivos de texto y una hoja de cálculo que contenga una lista de todos los nombres de archivo. Si le pide a sus anotadores que llenen las ranuras de la hoja de cálculo junto a cada nombre de archivo con la etiqueta, ¿qué sucede si están usando un programa que "útilmente" sugerirá opciones para autocompletar cada cuadro? Si está utilizando las etiquetas *positivo* , *negativo* o *neutral* , los dos últimos comienzan con "ne", y si un anotador se cansa o no presta atención, es posible que accidentalmente complete la etiqueta incorrecta. [Fi-](#)

**gura 5-1** muestra la facilidad con la que esto podría suceder en un editor de hojas de cálculo estándar. En una situación como esa, es posible que desee considerar el uso de un conjunto diferente de palabras, como *gustos*, *disgustos* e *indiferente*.

	A	B	
1	review001	positive	
2	review002	neutral	
3	review003	negative	
4	review004	positive	
5	review005	negative	
6	review006		
7	review007		
8	review008		

Figura 5-1. Una posible fuente de error en la anotación

Por supuesto, esto no significa que sea imposible completar una tarea utilizando una hoja de cálculo y clasificaciones que son un poco similares. En algunos casos, tales circunstancias son imposibles de evitar. Sin embargo, nunca es una mala idea estar atento a los lugares donde los errores pueden deslizarse fácilmente.

---

#### NOTA

Si bien no discutimos el escenario de anotación de reseñas de películas en **el Capítulo 4**, asumimos aquí que tenemos un esquema que contiene tres categorías. Sin embargo, esa no es de ninguna manera la única forma de enmarcar esta tarea y categorizar las reseñas de películas. En el Movie Review Corpus que viene con Natural Language Toolkit (NLTK), las reseñas se dividen solo en positivas y negativas (según las puntuaciones proporcionadas en las propias reseñas) y **RottenTomatoes.com** también utiliza una clasificación binaria. Por otro lado, **Metacritic.com** califica todo en una escala de 0 a 100.

Ambos sitios web brindan pautas de anotación para reseñas que no otorgan calificaciones numéricas preasignadas, y cada uno de esos sitios web tiene sus editores que asignan calificaciones en función de sus propios sistemas (**Metacritic.com ; RottenTomatoes.com**).

---

## Etiquetas Múltiples: Géneros Cinematográficos

A medida que sus tareas crecen en complejidad, hay más factores limitantes sobre cómo estructurar sus anotaciones. Por ejemplo, hay varias formas de abordar la tarea de etiquetar reseñas de películas que solo permiten una etiqueta por documento, pero ¿qué sucede si es posible que un documento tenga más de una etiqueta? En **el Capítulo 4**, comenzamos a discutir una especificación para una tarea que implica etiquetar resúmenes de películas con sus géneros asociados. Ampliemos ese ejemplo

ahora, para ver cómo podemos manejar tareas de anotación más complejas.

Si bien puede ser tentador decir simplemente: "Bueno, solo le daremos una sola etiqueta a cada película", intentar seguir esa pauta se vuelve difícil rápidamente. ¿Las comedias románticas se consideran romances o comedias? Podría agregar "comedia romántica" como una etiqueta de género, pero ¿creará una nueva etiqueta para cada película que cruce la línea de género? Tal tarea rápidamente se vuelve ridícula, simplemente debido a la cantidad de combinaciones posibles. Por lo tanto, defina sus géneros y permita que los anotadores coloquen tantas etiquetas como sea necesario en cada película (en el [Capítulo 6](#) discutiremos con más detalle los posibles enfoques de las pautas para tal tarea).

Entonces, ¿cómo se debe capturar esta información? De las opciones enumeradas para la tarea de revisión de películas, algunas de ellas se pueden descartar de inmediato. Hacer que sus anotadores cambien los nombres de los archivos para que contengan las etiquetas probablemente sea engorroso tanto para los anotadores como para usted: *Casablanca-drama.txt* es bastante fácil, pero *Spaceballs-sciencefiction\_comedy\_action\_parody.txt* sería molesto de crear para un anotador, y igualmente molesto para usted analizar en una forma más utilizable (especialmente si los errores de ortografía comienzan a colarse).

Mover archivos a carpetas debidamente etiquetadas también es más difícil con esta tarea; se tendría que crear una copia del archivo para cada etiqueta, y sería mucho más difícil recopilar información básica, como cuántas etiquetas, en promedio, se le dio a cada película. También sería mucho, mucho más difícil para los anotadores determinar si se perdieron una etiqueta.

En [la Figura 5-1](#), mostramos una hoja de cálculo de muestra con nombres de archivo y etiquetas positivas/negativas/neutrales en diferentes columnas, con una fila diferente para cada revisión. Si bien sería posible crear una hoja de cálculo configurada de la misma manera para dar a sus anotadores, no es difícil imaginar cuán propensa a errores sería ese tipo de entrada para una tarea con aún más opciones de categoría y más columnas potenciales por película. .

Entonces, ¿dónde nos deja eso? Si ninguna de las formas más sencillas de etiquetar datos está disponible, probablemente sea el momento de buscar herramientas de anotación y representaciones XML de datos de anotación.

En este caso, dado que la información que desea capturar son metadatos que son relevantes para todo el documento, probablemente no necesite preocuparse por las compensaciones de caracteres, por lo que puede tener etiquetas que se vean así:

```
<GenreXML>
  <FILM fid = "f1" title = "Cowboys and Aliens" file_name = "film01.txt" />
  <GÉNERO gid = "g1" filmid = "f01" label = "occidental" />
  <GÉNERO gid = "g2" filmid = "f01" label = "ciencia ficción" />
  <GÉNERO gid = "g3" filmid = "f01" etiqueta = "acción" />
</GENREXML>
```

Esta es una anotación muy simple, con una DTD o definición de tipo de documento igualmente simple [si no está seguro de cómo leer esta DTD, consulte la barra lateral ¿ [Qué es una DTD?](#) ]:

```
<!ENTIDAD nombre "GenreXML">

<!PELÍCULA DE ELEMENTO (#PCDATA) >
<!Identificador de identificación de la película ATTLIST >
<!ATTLIST PELÍCULA título CDATA>
<!ATTLIST FILM nombre_archivo CDATA >

<!GÉNERO DEL ELEMENTO (#PCDATA) >
<!ATTLIST GÉNERO id ID >
<!ATTLIST GENRE filmid CDATA >
<!ATTLIST GÉNERO etiqueta ( acción | aventura | clásico | ... ) >
```

Esta representación de la tarea de etiquetado de género no es la única forma de abordar el problema (en el [Capítulo 4](#) le mostramos una especificación ligeramente diferente para la misma tarea). Aquí tenemos dos elementos, `film` y `genre`, cada uno con un número de identificación y atributos relevantes; el `genre` elemento está vinculado a la película que representa por el `filmid` atributo.

---

#### ADVERTENCIA

No caiga en la trampa de pensar que existe una especificación verdadera para su tarea. Si encuentra que es más fácil estructurar sus datos de cierta manera, o agregar o eliminar elementos o atributos, ¡hágalo! No permita que sus especificaciones se interpongan en el camino de su objetivo.

---

Al tener el nombre del archivo almacenado en el XML para la lista de géneros, es posible mantener la anotación completamente separada del

texto del archivo que se está anotando. Sin embargo, claramente el `file_name` atributo no es obligatorio, y probablemente no desee que un anotador lo complete a mano. Pero es útil y sería fácil de generar automáticamente durante el procesamiento previo o posterior de los datos de anotación.

Dar a cada etiqueta un número de identificación (en lugar de solo las `FILM` etiquetas) puede no parecer muy importante en este momento, pero es un buen hábito porque hace que discutir y modificar los datos sea mucho más fácil, y también puede hacer que sea más fácil expandir su anotación. tarea más tarde si es necesario.

En este punto, es posible que se pregunte cómo toda esta información adicional lo ayudará con su tarea. Hay algunas razones por las que debería estar dispuesto a asumir estos gastos generales adicionales:

- Tener un elemento que contenga la información de la película permite que la anotación se mantenga en el mismo archivo que el resumen de la película o en otro lugar sin perder la pista de los datos.
- Mantener los datos en un formato estructurado le permite manipularlos más fácilmente más adelante. Hacer que la anotación tome la forma de XML bien formateado puede hacer que sea mucho más fácil de analizar más adelante.
- Ser capaz de crear una representación estructurada de su especificación ayuda a consolidar su tarea y puede mostrarle dónde están los problemas en la forma en que está pensando en su objetivo.
- Representar su especificación como una DTD (u otro formato) significa que puede usar herramientas de anotación para crear sus anotaciones. Esto puede ayudar a reducir la ortografía y otros errores de entrada del usuario.

**La Figura 5-2** muestra cómo se ve la tarea de anotación de géneros cinematográficos en el Entorno de anotación multipropósito (MAE), una herramienta de anotación que solo requiere un documento similar a DTD para configurarse y ejecutarse. Como puede ver, al tener las opciones de género proporcionadas en la DTD, un anotador solo tiene que crear una nueva instancia del `GENRE` elemento y seleccionar el atributo que desea de la lista.



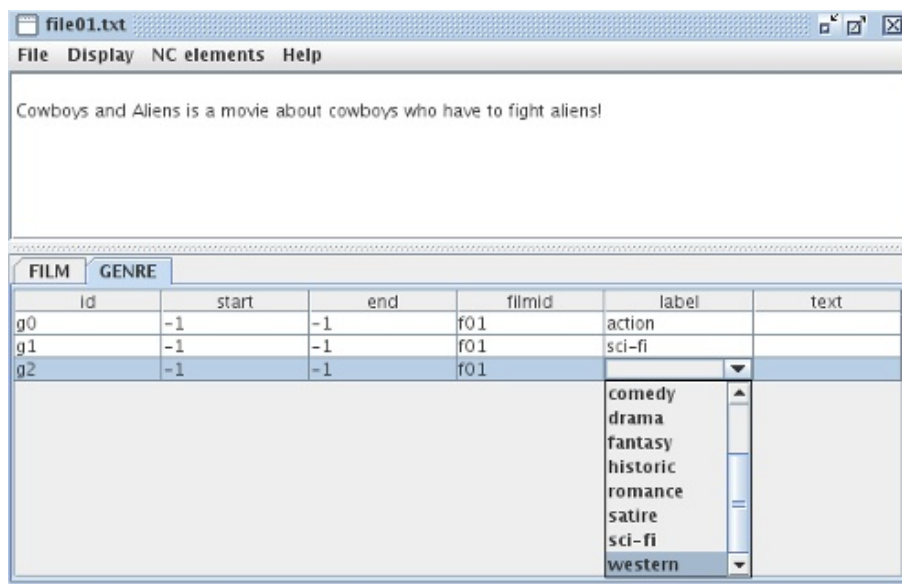


Figura 5-2. Anotación de género en MAE

El resultado de este proceso de anotación se vería así:

```
<FILM id="f0" start="-1" end="-1" text="" title="Cowboys and Aliens" />
<GÉNERO id="g0" start="-1" end="-1" text="" label="action" />
<GÉNERO id="g1" start="-1" end="-1" text="" label="ciencia ficción" />
<GÉNERO id="g2" start="-1" end="-1" text="" label="western" />
```

Aquí hay algunos elementos más que los especificados en la DTD que se muestra anteriormente: la mayoría de las herramientas requerirán que se cumplan ciertos parámetros para poder trabajar con una tarea, pero en la mayoría de los casos esos cambios son superficiales. En este caso, dado que MAE generalmente se usa para anotar partes del texto en lugar de crear metaetiquetas, se tuvo que cambiar la DTD para permitir que MAE creara GENRE y FILM no consumiera etiquetas. Es por eso que los elementos start y end se establecen en -1, para indicar que el alcance de la etiqueta no se limita a ciertos caracteres del texto. Notará que aquí, el filmid atributo en la GENRE etiqueta no está presente, y tampoco lo está el file\_name atributo en el FILM etiqueta. Si bien no sería descabellado pedirles a sus anotadores que asignen esa información ellos mismos, sería más fácil, además de más rápido y más preciso, hacerlo con un programa.

Si planea mantener la anotación separada en el mismo archivo que el texto que se está anotando, es posible que no necesite agregar la información del archivo a cada etiqueta. Sin embargo, los datos de las anotaciones pueden ser mucho más fáciles de analizar/manipular si no es necesario extraerlos del texto al que se refieren, por lo que mantener la información de la etiqueta en diferentes archivos que se refieran a los originales suele ser una buena práctica.



# Anotación de extensión de texto: entidades con nombre

Las tareas de clasificación de revisión e identificación de género son ejemplos de etiquetas de anotación que se refieren a la totalidad de un documento. Sin embargo, muchas tareas de anotación requieren un enfoque más detallado, donde las etiquetas se aplican a áreas específicas del texto, en lugar de todo a la vez. Ya analizamos muchos ejemplos de este tipo de tarea: etiquetado de parte del discurso (POS), reconocimiento de entidad con nombre (NE), las partes de identificación de tiempo y evento de TimeML, etc. Básicamente, cualquier proyecto de anotación que requiera que las secciones del texto reciban etiquetas distintas se incluye en esta categoría. Nos referiremos a esto como *anotación de extensión*, porque está anotando una extensión de texto en los datos que se pueden asociar con ubicaciones de caracteres.

En **el Capítulo 4** discutimos las diferencias entre la anotación independiente y en línea, y las extensiones de texto son donde las diferencias se vuelven importantes. Las etiquetas de tipo metadatos utilizadas para la tarea de clasificación de documentos podrían contener indicadores de inicio y finalización o podrían omitirlos; su presencia en el software de anotación fue un artefacto del propio software, más que una declaración de mejores prácticas. Sin embargo, con la anotación de separación, se requiere que los indicadores de ubicación estén presentes en cada etiqueta. Naturalmente, existen múltiples formas de almacenar esta información, como por ejemplo:

- anotación en línea
- Anotación separada por ubicación en una oración o párrafo
- Anotación separada por ubicación de carácter

En las siguientes secciones discutiremos las aplicaciones prácticas de cada uno de estos métodos, utilizando la anotación de entidad nombrada como caso de estudio.

Como discutimos anteriormente, la anotación NE se refiere a marcar lo que probablemente considere nombres propios: objetos en el mundo real que tienen designadores específicos, no solo etiquetas genéricas. Entonces, “El Empire State Building” es un NE, mientras que “el edificio de allá” no lo es. Por ahora, usaremos la siguiente especificación para describir la tarea NE:

```
<!ENTIDAD nombre "NamedEntityXML">
```

```
<!ELEMENTO NE (#DATOSPC) >
```

```
<!ATTLIST NE id ID >
```

```
<!ATTLIST NE tipo (persona | título | país | edificio | negocio | ...) >
```

```
<!ATTLIST NE nota CDATA>
```

## Anotación en línea

Si bien aún recomendamos enfáticamente no usar esta forma de almacenamiento de datos para su proyecto de anotación, el hecho es que es una forma común de almacenar datos. La frase "anotación en línea" se refiere a las etiquetas XML de anotación que están presentes en el texto que se está anotando y que rodean físicamente la extensión a la que se refiere la etiqueta, así:

```
<NE id="i0" type="building">La Casa del Estado de  
Massachusetts</NE> en <NE id="i1" type="city">Boston, MA</NE>  
alberga las oficinas de muchas figuras importantes del estado ,  
incluidos <NE id="i2" type="title">Gobernador</NE> <NE id="i3"  
type="person">Deval Patrick</NE> y los de <NE id="i4"  
type="organization">Tribunal General de Massachusetts</NE>.
```

Por lo menos, este formato de anotación es extremadamente difícil de leer. Pero más importante, cambia el formato del texto original. Si bien en este pequeño ejemplo puede que no haya nada especial en el formato del texto, la estructura física de otros documentos puede ser importante para un análisis posterior, y la anotación en línea dificulta su conservación o reconstrucción. Además, si esta anotación se fusionara más adelante, por ejemplo, con el etiquetado de POS, el dolor de cabeza de lograr que los dos conjuntos de etiquetas diferentes se superpongan podría ser enorme.

No todas las formas de anotación en línea están en formato XML. Hay otras formas de marcar los datos que están dentro del texto, como usar paréntesis para marcar grupos sintácticos, como se hizo en el siguiente ejemplo de Penn TreeBank II, tomado de "The Penn TreeBank: anotación de la estructura del argumento predicado" ( [Marcus et al. 1994](#) ):

```
(S (NP-SUBJ I  
  (VP considerar  
    (S (NP-SUBJ Kris)  
      (NP-PRD un tonto))))))
```

Todavía hay muchos programas que proporcionan salida en este formato o uno similar (elStanford Dependency Parser es un ejemplo), y si desea utilizar herramientas que hagan esto, es posible que deba encontrar una manera de convertir la información en este formato en una anotación independiente para que sea lo más portátil posible para otras aplicaciones.

Por supuesto, hay algunos beneficios para la anotación en línea: se vuelve innecesario realizar un seguimiento especial de la ubicación de las etiquetas o el texto que las etiquetas rodean, porque esas cosas son inseparables. Aún así, estos beneficios son bastante miopes y recomendamos encarecidamente no utilizar este paradigma para la anotación.

Otro tipo de anotación en línea se ve comúnmente en el etiquetado de POS u otras tareas en las que una etiqueta se asigna a una sola palabra (en lugar de abarcar muchas palabras). De hecho, ya vio un ejemplo en el [\*\*Capítulo 1\*\*](#), en la discusión del Penn TreeBank.

“/” De/EN el/DT inicio/NN ./, tomó/PRP/VBD un/DT hombre/NN con/EN  
extraordinarias/JJ cualidades/NNS para/PARA triunfar/VB en/EN  
México/NNP ./, “/” dice/VBZ Kimihide/NNP Takimura/NNP ./,  
presidente/NN de/IN Mitsui/grupo NNS/NN 's/POS Kensetsu/NNP  
Engineering/NNP Inc./unidad NNP/NN ./.

Aquí, cada etiqueta POS se agrega como un sufijo directamente a la palabra a la que se refiere, sin ninguna etiqueta XML que separe la extensión de su etiqueta. Esta forma de anotación no solo dificulta la lectura de los datos, sino que también cambia la composición de las propias palabras. Considere cómo "group's" se convierte en "group/NN 's/POS": el posesivo "'s" se ha separado de "group", lo que ahora dificulta aún más la reconstrucción del texto original. ¡O imagine tratar de reconciliar una anotación como esta con el ejemplo NE en el ejemplo anterior! No sería imposible, pero sin duda podría causar dolores de cabeza.

Si bien generalmente tampoco recomendamos usar este formato, muchos etiquetadores de POS existentes y otras herramientas se escribieron originalmente para proporcionar resultados de esta manera, por lo que es algo que debe tener en cuenta, ya que es posible que deba realinear el texto original con el nuevas etiquetas POS.

Por supuesto, no estamos sugiriendo que nunca deba usar herramientas que generen información en formatos que no sean algunas anotaciones independientes variantes. Muchas de estas herramientas son extremadamente útiles y proporcionan resultados muy precisos. Sin embargo, debe tener en cuenta los problemas que pueden surgir al intentar utilizarlos.

---

Otro problema con este formato de anotación es que si se aplica a la tarea NE, existe el problema inmediato de que la tarea NE requiere que una sola etiqueta se aplique a más de una palabra *al mismo tiempo*. Hay una distinción importante entre aplicar la misma etiqueta más de una vez en un documento (ya que hay más de una NN etiqueta en el ejemplo de Penn TreeBank) y aplicar una etiqueta en un lapso de palabras. Agrupar un conjunto de palabras usando una sola etiqueta le dice al lector algo acerca de ese grupo que no le dice al aplicar la misma etiqueta a cada palabra individualmente. Considere estos dos ejemplos:

<NE id="i0"	El/Edificio_NE
type="building">La Casa del	Massachusetts/Edificio_NE
Estado de	Estado/Edificio_NE Casa/Edificio_NE
Massachusetts</NE> en <NE	en Boston/Ciudad_NE ,/Ciudad_NE
id="i1" type="city">Boston,	MA/Ciudad_NE ...
MA</NE> ...	

En el ejemplo de la izquierda, está claro que la frase "La Casa del Estado de Massachusetts" es una unidad en lo que respecta a la anotación: la NE etiqueta se aplica a todo el grupo. Por otro lado, en el ejemplo de la derecha, la misma etiqueta se aplica individualmente a cada token, lo que hace que sea mucho más difícil determinar si cada token es un NE por sí mismo o si existe una conexión entre ellos. De hecho, terminamos etiquetando algunas fichas con la etiqueta incorrecta. Tenga en cuenta que el estado "MA" debe identificarse como "/NE\_city" para que el tramo se reconozca como una ciudad.

### **Anotación de separación por tokens**

Un método que a veces se utiliza para la anotación independiente es la *tokenización* (es decir, la separación) de la entrada de texto y la asignación de un número a cada token. El proceso de tokenización generalmente se basa en espacios en blanco y puntuación, aunque el proceso específico puede variar según el programa (por ejemplo, algunos programas dividirán "'s" o "n't" de "Meg's" y "don't", y otros lo harán). El texto en el

ejemplo de anotación adjunto se ha tokenizado: cada palabra y signo de puntuación se ha separado.

Tomando el texto anterior como ejemplo, hay algunas formas diferentes de identificar el texto asignando números a las fichas. Una forma es simplemente numerar cada ficha en orden, comenzando en 1 (o 0, si lo prefiere) y continuando hasta que no queden más fichas, como se muestra en la Tabla **5-1** .

*Tabla 5-1. Etiquetado de palabras por token*

SIMBÓLICO	TOKEN_ID
“	1
De	2
el	3
comienzo	4
,	5
...	...
unidad	31
.	32

Estos datos podrían almacenarse en un archivo separado por tabulaciones o en una hoja de cálculo, ya que es necesario mantener las ID asociadas con cada token. Otra forma es asignar números a cada oración e identificar cada ficha por número de oración y su lugar en esa oración, como se muestra en la **Tabla 5-2** .

Tabla 5-2. Etiquetado de palabras por oración y token

SIMBÓLICO	ENVIADO_ID	TOKEN_ID
“	1	1
De	1	2
el	1	3
comienzo	1	4
,	1	5
...		
unidad	1	31
.	1	32
Entonces	2	1
...		

Naturalmente, se podrían agregar más características de identificación, como el número de párrafo, el número de documento, etc. La ventaja de tener información adicional (como el número de oración) para identificar tokens es que esta información se puede usar más tarde para ayudar a definir características para los algoritmos de aprendizaje automático (ML) (mientras que el número de oración se puede inferir nuevamente más adelante, si se sabe que es importante, entonces es más fácil tener esa información por adelantado).

Los datos de anotación que usan este formato podrían parecerse a **la Tabla 5-3**.

Tabla 5-3. Anotación de POS en texto tokenizado

POS_TAG	ENVIADO_ID	TOKEN_ID
“	1	1
EN	1	2
DT	1	3
NN	1	4
...		

Hay algunas ventajas en el uso de este formato: debido a que la anotación se elimina del texto, no es necesario preocuparse por la superposición de etiquetas al intentar fusionar anotaciones realizadas en los mismos datos. Además, esta forma de anotación sería relativamente fácil de configurar con un programa tokenizador y cualquier texto que desee darle.

Sin embargo, también hay algunos problemas con esta forma de anotación. Como puede ver, debido a que el texto está dividido en espacios en blanco y puntuación, el formato original de los datos no se puede recuperar, por lo que se ha violado la máxima de "no dañar" los datos. Si la estructura del documento en el que apareció este texto más tarde se volvió importante al crear funciones para un clasificador, podría ser difícil fusionar esta anotación con el formato de texto original.

---

**NOTA**

Es posible usar anotaciones basadas en tokens sin dañar los datos, aunque sería necesario ejecutar el tokenizador cada vez que la anotación necesitara emparejarse con el texto, y siempre se tendría que usar el mismo tokenizador. Esta es la forma sugerida para tratar con la anotación de separación basada en tokens.

---

Además, este formato tiene un problema similar al de la anotación adjunta, ya que parece suponer que cada etiqueta se aplica a un solo token. Si bien no es imposible aplicar una etiqueta a un conjunto de tokens, la sobrecarga se vuelve mayor. Considere nuevamente nuestro ejemplo de NE, esta vez tokenizado, como se muestra en [la Tabla 5-4](#).



Tabla 5-4. Etiquetas de corpus basadas en tokens

SIMBÓLICO	ENVIADO_ID	TOKEN_ID
El	1	1
Massachusetts	1	2
Estado	1	3
Casa	1	4
en	1	5
Bostón	1	6
,	1	7
MAMÁ	1	8
casas	1	9
...		

**La tabla 5-5** muestra cómo aplicaríamos una etiqueta que abarca varios tokens.

Tabla 5-5. Ejemplo de anotación de separación utilizando identificadores de token

ETIQUETA	START_SENT_ID	START_TOKEN_ID	END_SENT_ID	END_TOKEN_ID
NE_building	1	1	1	4
NE_city	1	6	1	8

El otro defecto de este método es que no permite anotar fácilmente partes de una palabra. Los proyectos de anotación centrados en morfemas o raíces verbales requerirían anotar tokens parciales, lo que sería difícil con este método. No es imposible de hacer: se podría usar otro conjunto de atributos para cada token para indicar qué caracteres del token se etiquetan. Sin embargo, llegados a ese punto, también se podría pasar a la ano-

tación independiente basada en caracteres, de la que hablaremos en la siguiente sección.

## **Anotación de separación por ubicación de caracteres**

El uso de ubicaciones de caracteres para definir a qué parte del documento se aplica una etiqueta es una forma confiable de generar una anotación independiente que se puede usar en diferentes sistemas. Las anotaciones basadas en caracteres utilizan la información de desplazamiento de caracteres para colocar etiquetas en un documento, como esta:

```
La Casa del Estado de Massachusetts en Boston, MA alberga las oficinas
de muchas figuras estatales importantes, incluido el gobernador Deval Patrick
y los del Tribunal General de Massachusetts.
```

```
<NE id="N0" start="5" end="31" text="Casa del Estado de Massachusetts"
    tipo="edificio" />
<NE id="N1" start="35" end="45" text="Boston, MA" type="city" />
<NE id="N2" start="109" end="117" text="Gobernador" type="title" />
<NE id="N3" start="118" end="131" text="Deval Patrick" type="persona" />
<NE id="N4" start="150" end="177" text="Tribunal General de Massachusetts"
    tipo="organización" />
```

A primera vista, es difícil ver los beneficios de este formato para la anotación: los números de inicio y final no significan mucho para alguien que solo mira las etiquetas, y las etiquetas están tan alejadas del texto que casi no tienen relación. Sin embargo, esta distancia es precisamente la razón por la que la anotación de separación es importante, incluso necesaria. Al separar las etiquetas del texto, es posible tener muchas anotaciones diferentes que apuntan al mismo documento sin interferir entre sí y, lo que es más importante, sin cambiar el texto original del documento.

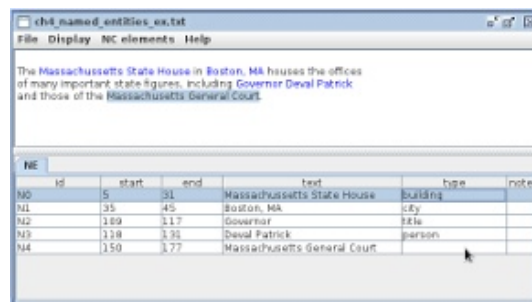
En cuanto a los números de inicio y final, si bien es difícil para un ser humano determinar a qué se refieren, es muy fácil para las computadoras contar las compensaciones para realizar esa tarea. Y cuanto más fácil sea para la computadora encontrar las partes importantes del texto, más fácil será usar ese texto y la anotación para el aprendizaje automático más adelante.

#### NOTA

El uso `start` de y `end` como nombres de atributo para indicar dónde deben colocarse las etiquetas en el texto es una convención que usamos aquí, pero no es un estándar en la anotación: diferentes herramientas y sistemas de anotación usarán diferentes términos para esta información. Del mismo modo, el `text` atributo tampoco tiene ningún significado especial. El nombre de los atributos no es importante; lo importante es la información que tienen.

Técnicamente, todo lo que se necesita para que estas etiquetas funcionen son las ubicaciones de desplazamiento inicial y final y los atributos de la etiqueta; aquí, las etiquetas también contienen el texto al que se aplica la etiqueta, porque hace que la anotación sea más fácil de evaluar. Incluso si esa información no estuviera allí, la etiqueta seguiría siendo funcional.

**La Figura 5-3** muestra cómo sería crear esta anotación en una herramienta de anotación.



NE	Id	start	end	text	type	note
N0	5	31		Massachusetts State House	Building	
N1	35	45		Boston, MA	City	
N2	109	117		Governor	Role	
N3	118	136		Deval Patrick	Person	
N4	150	177		Massachusetts General Court		

Figura 5-3. anotación NE

Naturalmente, se necesita algo de preparación para que la anotación separada funcione bien. Para empezar, es importante decidir al principio del proceso qué codificación de caracteres usará para su corpus y mantenerla durante todo el proceso de anotación. La codificación de caracteres que elija determinará cómo las diferentes computadoras y programas cuentan dónde están los caracteres en sus datos, y cambiar las codificaciones a la mitad puede causar que se pierda mucho trabajo. Recomendamos utilizar la codificación UTF-8 para sus datos.

#### ADVERTENCIA

Los problemas de codificación pueden causar muchos dolores de cabeza, especialmente si sus datos se transferirán entre computadoras que usan diferentes sistemas operativos. El uso de Windows puede hacer que esto sea particularmente difícil, ya que parece que Windows no usa la codificación UTF-8 de forma predeterminada, mientras que la mayoría de los otros sistemas operativos (Mac y la mayoría de las versiones de Unix/Linux que conocemos) sí lo hacen. No es imposible usar UTF-8 en Windows, pero requiere un pequeño esfuerzo adicional.

## Anotación de extensión vinculada: roles semánticos

A veces, en las tareas de anotación, es necesario representar la conexión entre dos extensiones etiquetadas diferentes. Por ejemplo, en la anotación temporal, no es suficiente anotar “Lunes” y “corrió” en la oración “Juan corrió el lunes”; para representar completamente la información presentada en la oración, también debemos indicar que existe una conexión entre el día y el evento. Esto se hace usando etiquetas de relación, también llamadas etiquetas de enlace

Veamos de nuevo nuestra oración de ejemplo sobre Boston. Si quisiéramos agregar información de ubicación a esta anotación, querríamos una forma de indicar que existe una relación entre los lugares. Podríamos hacerlo agregando una etiqueta a nuestra DTD que se vería así:

```
<!ELEMENT L_LINK VACÍO>
<!ATTLIST L-LINK de ID IDREF >
<!ATTLIST L-LINK a ID IDREF>
<!ATTLIST L-LINK relación (interior | exterior | mismo | otro ) >
```

Obviamente, este es un conjunto muy limitado de relaciones de ubicación, pero funcionará por ahora. ¿Cómo se aplicaría esto a la anotación que ya tenemos?

Aquí es donde las identificaciones de etiqueta que mencionamos en **Múltiples etiquetas: géneros cinematográficos** se vuelven muy importantes. Dado que las etiquetas de enlace no se refieren directamente a extensiones en el texto, deben representar la conexión entre dos objetos anotados. La forma más común de representar esta información es usar los números de identificación de las etiquetas de extensión para anclar los enlaces. Esta nueva información se verá así:

La Casa del Estado de Massachusetts en Boston, MA alberga las oficinas de muchas figuras estatales importantes, incluido el gobernador Deval Patrick y los del Tribunal General de Massachusetts.

```
<NE id="N0" start="5" end="31" text="Casa del Estado de Massachusetts"
    tipo="edificio" />
<NE id="N1" start="35" end="45" text="Boston, MA" type="city" />
<NE id="N2" start="109" end="117" text="Gobernador" type="title" />
<NE id="N3" start="118" end="131" text="Deval Patrick" type="persona" />
<NE id="N4" start="150" end="177" text="Tribunal General de Massachusetts"
    tipo="organización" />
```

```
<L-LINK id="L0" fromID="N0" toID="N1" relación="dentro" />  
<L-LINK id="L0" fromID="N4" toID="N0" relación="dentro" />
```

Al referirnos a los ID de las NE etiquetas, podemos codificar fácilmente información sobre las relaciones entre ellas. Y debido a que las L-LINK etiquetas también tienen números de identificación, también es posible crear conexiones entre ellas; tal vez un nivel más alto de anotación podría indicar que dos L-LINK s representan la misma información de ubicación, lo que podría ser útil para un proyecto diferente.

---

#### NOTA

Una vez más, los nombres de los atributos aquí no son particularmente importantes. Usamos `fromID` y `toID` como nombres para los enlaces anclados, porque eso es lo que hace la herramienta de anotación MAE, pero otro software usa convenciones diferentes. La intención, sin embargo, es la misma.

---

## Las normas ISO y usted

En [Normas ISO](#) discutimos la LAF (Linguistic Annotation Framework) estándar para representar datos anotados. Puede que haya sonado bastante formal, ¡pero no te preocupes! Si sigue nuestras recomendaciones de este libro y utiliza una anotación separada basada en XML, es probable que su estructura de anotación ya sea compatible con LAF y que solo necesite convertirla al formato de volcado LAF. También tenga en cuenta que LAF es una excelente base para los investigadores lingüísticos que desean compartir sus datos, pero si su anotación solo está destinada a usted o es propiedad de una empresa, es posible que esto no sea algo de lo que deba preocuparse. .

## Resumen

En este capítulo discutimos algunos de los diferentes métodos para representar anotaciones para una tarea de anotación de corpus. En particular, notamos lo siguiente:

- Las anotaciones se pueden almacenar de muchas maneras diferentes, pero es importante elegir un formato que sea flexible y fácil de cambiar más adelante si es necesario. Recomendamos formatos independientes basados en XML.
- En algunos casos, como las tareas de clasificación de documentos de una sola etiqueta, hay muchas formas de almacenar datos de anotación, pero estas técnicas son esencialmente isomorfas. En tales ca-

sos, elija qué método usar considerando cómo planea usar los datos y qué métodos funcionan mejor para sus anotadores.

- Para la mayoría de las tareas de anotación, como aquellas que requieren múltiples etiquetas en un documento, y especialmente aquellas que requieren anotación y vinculación de extensión, será útil tener un software de anotación para que lo utilicen sus anotadores. Consulte **el Apéndice B** para obtener una lista del software disponible.
- La anotación de extensión puede adoptar muchas formas, pero la anotación separada basada en caracteres es el formato que facilitará la realización de los cambios necesarios en las anotaciones más adelante y también facilitará la combinación con otras anotaciones.
- Si decide utilizar una anotación independiente basada en caracteres, tenga cuidado con las codificaciones que utiliza para sus datos, especialmente cuando crea el corpus en primer lugar. Los diferentes lenguajes de programación y sistemas operativos tienen diferentes configuraciones predeterminadas para la codificación de caracteres, y es vital que use un formato que funcione para todos sus anotadores (o al menos esté dispuesto a dictar qué recursos usan sus anotadores).
- El uso de estándares de la industria como XML y estándares de anotación como LAF para sus anotaciones le facilitará mucho la interacción con otros corpus anotados y le facilitará compartir su propio trabajo.

## Capítulo 6. Anotacion y adjudicacion

Ahora que tiene un corpus y un modelo, es hora de comenzar a observar el proceso de anotación real: la "A" en el ciclo MATTER. Aquí es donde define el método por el cual su modelo se aplica a sus textos, tanto en teoría (cómo se describe su tarea a los anotadores) como en la práctica (qué software y otras herramientas se utilizan para crear las anotaciones). Una parte crítica de esta etapa es la adjudicación, donde toma el trabajo de sus anotadores y lo usa para crear el corpus *estándar de oro* que usará para el aprendizaje automático. En este capítulo responderemos a las siguientes preguntas:

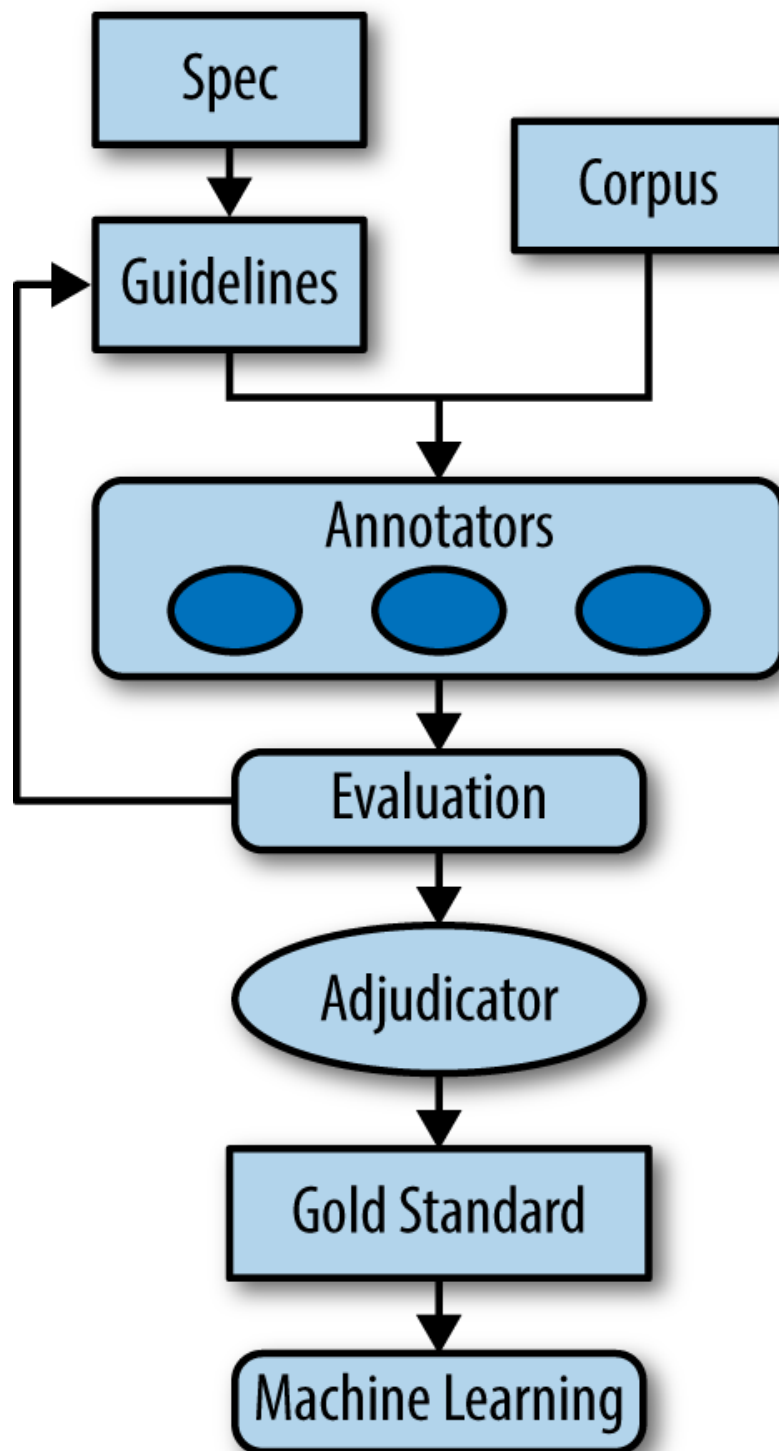
- ¿Cuáles son los componentes de una tarea de anotación?
- ¿Cuál es la diferencia entre una especificación de modelo y las pautas de anotación?
- ¿Cómo crea pautas que se ajusten a su tarea?
- ¿Qué herramienta de anotación debe usar para su tarea de anotación?
- ¿Qué habilidades necesitan sus anotadores para crear sus anotaciones?
- ¿Cómo puede saber (cualitativamente) si sus pautas de anotación son buenas para su tarea?
- ¿Qué implica adjudicar las anotaciones?

### La infraestructura de un proyecto de anotación

Es mucho más fácil escribir pautas de anotación cuando comprende cómo se ejecutan normalmente los proyectos de anotación, por lo que antes de entrar en los detalles de la redacción de pautas, repasaremos algunas formas diferentes en las que puede estructurar su esfuerzo de anotación.

Actualmente, lo que llamaríamos el enfoque “tradicional” es así. Una vez que se desarrolla un esquema y se recopila un corpus, un investigador redacta pautas, encuentra anotadores y distribuye las pautas y el corpus a los anotadores, quienes se van, anotan y luego regresan con un corpus marcado. Luego, el investigador recopila los datos de cada uno de los anotadores y calcula las puntuaciones del acuerdo entre anotadores (IAA). Si estos puntajes son bajos, se revisan las pautas (ya veces el modelo) y se rehace la anotación. Si los puntajes son buenos, la adjudicación se realiza sobre los datos para crear un estándar de oro, que luego se usa para entrenar y probar un algoritmo de aprendizaje automático (ML). [La Figura 6-1](#) muestra visualmente el proceso de anotación.





*Figura 6-1. El proceso de anotación*

Entonces, ¿por qué los investigadores usan este método distribuido de anotación, en lugar de hacerlo ellos mismos? Hay dos razones principales. Primero, existe una preocupación puramente práctica: la mayoría de las tareas de anotación ahora se realizan con el objetivo final de entrenar un sistema ML, y para hacerlo, debe haber suficientes datos para entrenar el algoritmo. Dado que para la mayoría de las tareas esto requiere cientos o incluso miles de artículos o ejemplos anotados, sería completamente poco práctico que solo una o dos personas crearan el corpus requerido.

La segunda razón para la anotación distribuida es que al hacer que otras personas realicen la mayor parte del etiquetado y luego usar esos datos

para calcular las puntuaciones de concordancia, el investigador tiene una idea de si la tarea está lo suficientemente bien definida para proporcionar datos consistentes en caso de un corpus más grande. ser necesario. Si hay poco acuerdo entre los anotadores sobre cómo se debe usar cada etiqueta y atributo, puede ser una señal de que las pautas simplemente no están bien escritas, pero también puede ser una señal de que la tarea en sí tiene fallas o va en contra de los objetivos. forma en que la gente generalmente entiende el lenguaje.

Antes de que las computadoras fueran tan rápidas como lo son ahora, los investigadores que querían examinar los datos del idioma lo hacían a mano y, por lo general, solos o con la ayuda de un estudiante en el campo. Sin embargo, los corpus anotados anteriores a la década de 1980 tendían a ser relativamente pequeños, simplemente por la imposibilidad de crear y catalogar una gran cantidad de datos. Si bien se estaban reuniendo grandes corpus en las décadas de 1970 y 1980 (como el Brown Corpus), no fue hasta más tarde que los grandes esfuerzos de anotación basados en computadora se volvieron comunes.

Sin embargo, incluso con conjuntos de datos más grandes cada vez más disponibles, a veces todavía es necesario volver a una tarea existente y crear un corpus anotado más grande para ella. Por eso es importante tener pautas bien definidas y buenos puntajes IAA; si una tarea está clara, entonces puede ser reproducida más tarde por otros grupos de personas. En el **Capítulo 4** discutimos la interoperabilidad y la reproducibilidad en su aplicación a los formatos de datos, pero el concepto también se aplica a la tarea de anotación.

*Colaboración colectiva* es otro enfoque que se utiliza con más frecuencia en la comunidad de anotaciones. Esencialmente, en lugar de pedir a una pequeña cantidad de anotadores que etiqueten una gran cantidad de extensiones, la tarea se divide en una gran cantidad de tareas más pequeñas y se le pide a una gran cantidad de anotadores que etiqueten solo algunos ejemplos cada una. Una plataforma popular de crowdsourcing es **Mechanical Turk** (MTurk) de Amazon, un recurso donde las personas que tienen tareas que requieren inteligencia humana para realizarlas pueden realizar solicitudes que luego son atendidas por personas de todo el país y de todo el mundo. Estos HIT(tareas de inteligencia humana) generalmente están formateadas para ser muy rápidas y relativamente fáciles de realizar, por lo que a los trabajadores generalmente se les paga en centavos en lugar de dólares. Debido a la disponibilidad de anotadores y lo económico que es cada HIT, muchos investigadores han considerado utilizar el crowdsourcing como recurso para la anotación.

En general, MTurk solo será útil para tareas que no requieran ningún conocimiento lingüístico especial o de otro dominio. Además, debido a la naturaleza restringida de la interfaz HIT, los trabajadores no pueden recibir páginas y páginas de pautas de anotación; aparte de la impracticabilidad del diseño, generalmente no vale la pena el tiempo de un trabajador para leer tantas instrucciones por solo unos pocos centavos. . Entonces, si tiene una tarea de anotación que puede dividirse en pasos simples y explicarse fácilmente en unas pocas oraciones, entonces MTurk puede ser un buen recurso para que lo investigue. Debido a que tales tareas deben tener un formato diferente y dividirse en partes más pequeñas, no vamos a entrar en los detalles de la creación de buenos HIT de anotación en este libro, sino en el Apéndice **E**, tiene una lista de recursos relacionados con MTurk que pueden resultarle útiles si decide utilizar MTurk.

---

## Especificación Versus Directrices

En los capítulos anteriores, exploramos cómo se crean los modelos y analizamos diferentes formas de definir su tarea en términos de etiquetas y XML. Estas etiquetas encapsulan lo que su esquema es capaz de capturar: son una abstracción de lo que desea que representen sus datos. Las *pautas de anotación* representan la aplicación del esquema a los datos: son la guía sobre cómo se usa el esquema en sus datos.

---

### NOTA

Si le ayuda, piense en la distinción entre el esquema y las pautas como la diferencia entre una lista de ingredientes y las instrucciones reales para hornear. Las pautas determinan cómo se mezclan los ingredientes (el esquema: etiquetas, atributos y datos) y qué hacen cuando terminas. En la repostería, se pueden usar los mismos ingredientes para hacer pan o pasteles, y en la anotación, a veces se puede usar la misma especificación para diferentes proyectos de anotación.

---

La distinción entre esquema y directrices es importante, porque es posible utilizar un solo esquema para diferentes tareas de anotación, o para

la misma tarea básica pero con datos diferentes a los que se han utilizado en el pasado. Si ha decidido usar o modificar un esquema existente para su tarea de anotación, probablemente deba prestar especial atención a esta distinción, ya que es posible que las pautas existentes para su tarea no se ajusten a su objetivo o corpus específico.

Por ejemplo, las pautas de anotación originales de TimeML se escribieron para aplicarse a artículos de periódicos, lo que dio lugar a ejemplos y notas específicos en las pautas que no se aplicaban a otros géneros. Específicamente, cuando TimeML se aplicó a los datos médicos, se realizaron algunos cambios en el esquema, pero en su mayor parte, fueron las pautas de anotación las que debían modificarse para brindar ejemplos relevantes a los anotadores. Si bien puede sentir que es razonable esperar que un comentarista pueda extrapolar los eventos y las relaciones de "El presidente tiene la esperanza de que se pueda llegar a la reconciliación" a "El paciente tiene fiebre y puede experimentar náuseas", es mejor proporcionar información relevante para el género. Ejemplos que confiar demasiado en las intuiciones de sus anotadores.

Otro ejemplo de una pauta que cambia el uso de una especificación sería tener una tarea que limite la aplicación de la especificación. Considere las relaciones de entidad nombrada (NE) que analizamos en [Agregar entidades nombradas](#). Estos conjuntos de relaciones se derivaron de una especificación existente relacionada con *las etiquetas de roles semánticos*. Sin embargo, en lugar de usar todo el conjunto de relaciones semánticas de roles definidas en esa tarea, usamos solo un subconjunto de las que se relacionan con las películas y los roles que desempeñan las personas en su creación.

Finalmente, puede encontrarse en una situación en la que tiene una especificación que define su tarea, pero considera necesario dividir la tarea en diferentes pasos. A veces, esto es útil si tiene un modelo complicado y desea maximizar la precisión realizando anotaciones en solo unas pocas etiquetas a la vez, luego adjudicando esas etiquetas y luego pasando al siguiente conjunto de etiquetas o relaciones. También puede encontrar este enfoque necesario si su tarea requiere un conocimiento general del idioma y un conocimiento más específico que requeriría un experto en el dominio para proporcionar. En esos casos, puede tener dos tareas realizadas simultáneamente por diferentes grupos de personas en el mismo corpus, y luego fusionar las anotaciones más tarde para propósitos de ML.

## **Esté preparado para revisar**

Si bien es cierto que todo el proceso de MATERIA es un ciclo, es importante recordar que hay un ciclo más pequeño de revisión entre las etapas de Modelo y Anotación. Discutimos en [el Capítulo 4](#) cómo puede encontrar necesario reformular la especificación a medida que desarrolla su modelo, pero es muy probable que una vez que comience a escribir sus

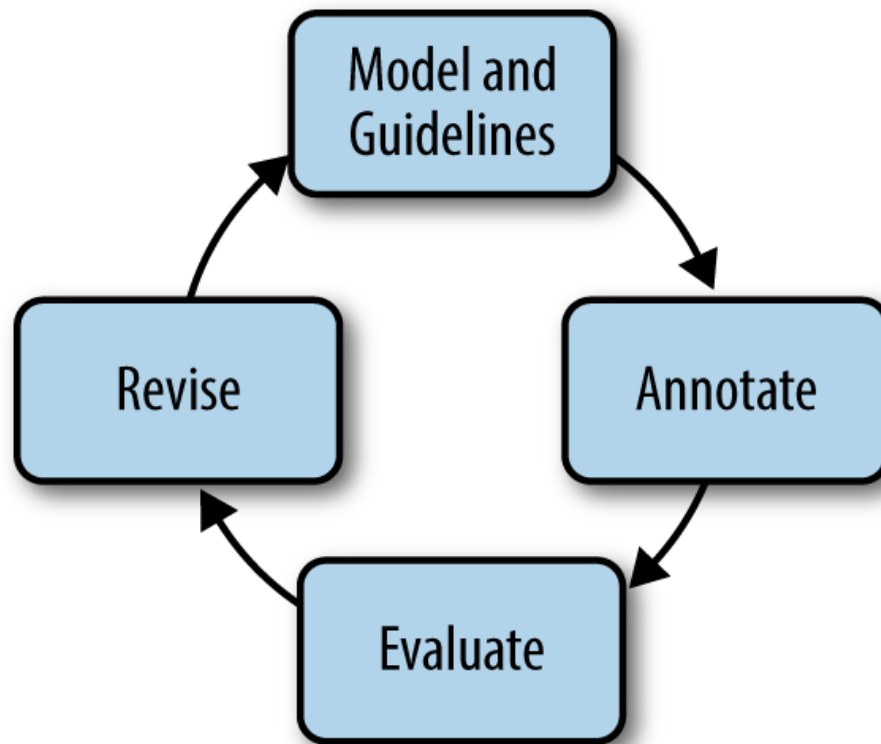
pautas de anotación, encontrará rápidamente lugares donde sus datos y especificaciones no coinciden. .

---

**NOTA**

La revisión de la especificación mientras se trabaja en las pautas y se realizan las anotaciones ocurre con tanta frecuencia que nos referimos a él como el ciclo MAMA: Modelar-Anotar-Modelar-Anotar (el paso Anotar incluye escribir las pautas). Consulte [la Figura 6-2](#) .

---



*Figura 6-2. El funcionamiento interno del ciclo MAMA*

Una vez que las pautas estén escritas y entregadas a sus anotadores, los anotadores mismos probablemente tendrán preguntas sobre las especificaciones y las pautas que pueden conducir a más revisiones. ¡Esto es completamente normal! Para tener una tarea de anotación que pueda reutilizarse en el futuro, es importante tomarse el tiempo para asegurarse de que las especificaciones y las pautas sean claras y estén bien escritas.

## **Preparando sus datos para la anotación**

La preparación de datos se divide en dos tareas principales: el acto más teórico de decidir cómo presentar sus documentos a sus anotadores y el asunto muy práctico de forzar esos documentos en una forma que funcionará con la herramienta de anotación que elija. Una discusión sobre las herramientas de anotación tendrá lugar más adelante en este capítulo, así que por ahora, solo nos centraremos en decidir qué información presentar a sus anotadores.

## Metadatos

Decidir qué información dar a sus anotadores puede ser más complicado de lo que parece. Tiene los archivos de texto que desea anotar, pero ¿cuánto desea que sepan sus anotadores sobre el origen de esos archivos? Por ejemplo, ¿deberían saber de dónde provienen los archivos y quién los produjo?

Esto puede parecer un problema trivial, pero la información sobre un documento puede influir en los anotadores de formas a veces inesperadas. La anotación de opinión en particular puede ser susceptible a este problema. Un ejemplo obvio sería, al examinar la polaridad de las reseñas de películas, informar a los anotadores de la calificación de estrellas que el crítico le dio a la película. Si desea que un anotador busque tanto los adjetivos positivos como los negativos, informar al anotador de que el crítico le dio a la película 10 estrellas puede prepararlo para buscar solo los adjetivos positivos.

Del mismo modo, si los anotadores están examinando artículos de periódicos o de Wikipedia en busca de hechos, informándoles de la fuente del artículo, o si la página de Wikipedia fue marcada por ser parcial, podría influir en los anotadores, particularmente si una fuente es, para ellos, asociados con informes sesgados, o si tienden a creer que las banderas de Wikipedia son precisas.

Como otro ejemplo, digamos que está trabajando en la creación de una transcripción fonética de datos grabados; informar a sus anotadores del lugar de nacimiento de un hablante podría potencialmente sesgarlos para escuchar pronunciaciones de una manera particular, especialmente si son nuevos en la transcripción fonética.

### PRECAUCIÓN

El sesgo puede aparecer de formas inesperadas; es útil probar algunas variaciones en la presentación de datos para eliminar diferentes factores de sesgo.

## Datos preprocesados

Otra consideración en la presentación de datos es proporcionar a los anotadores datos que ya tienen información marcada.

Sin embargo, hay dos cosas a considerar antes de hacer eso:

- ¿Esta información es necesaria para un anotador humano o sería más útil para un algoritmo?

- ¿Esta información será correcta o es algo que el anotador tendrá que corregir/examinar en busca de errores?

La parte del discurso (POS) es un excelente ejemplo de información que sería útil para una máquina, pero que podría no ser necesario proporcionar a un anotador (especialmente si el anotador es un hablante nativo del idioma que está evaluando). ). Presentar a un anotador demasiada información puede hacer que la tarea sea confusa y podría reducir la precisión sin aumentar necesariamente la informatividad.

Si su información preprocesada es correcta o no, tendrá un gran impacto en su proyecto de anotación. Ciertamente, ahorra tiempo hacer parte de su anotación automáticamente si existen las herramientas, pero a la gente en general le gustan las instrucciones simples, por lo que "Etiquetar todas las Entidades/horas/eventos con nombre, luego crear enlaces entre ellos" podría ser una opción más fácil. regla a seguir (especialmente porque podría realizarse como dos tareas diferentes) que "Hemos marcado la mayoría de las entidades/tiempos/eventos para usted, pero es posible que no sean correctos, por lo que nos gustaría que corrigiera cualquier error mientras creando enlaces."

Si las anotaciones proporcionadas son de mala calidad, pedirles a los anotadores que corrijan las anotaciones parciales puede provocar que no se solucionen los errores, especialmente cuando se combinan con otras instrucciones al mismo tiempo. Dividir la anotación en tareas relacionadas, como corregir las etiquetas generadas y luego agregar más etiquetas o crear vínculos entre las extensiones corregidas, es probablemente la mejor manera de lidiar con esto. Consume más tiempo, pero dará como resultado anotaciones de mayor calidad.

### **Dividir los archivos para la anotación**

Ya le contamos cómo necesitará dividir sus archivos en conjuntos de datos de entrenamiento y prueba, pero también querrá reservar algunos documentos para probar sus pautas de anotación. A medida que avanza en el ciclo de MAMA, encontrará que (especialmente al principio) no les llevará mucho tiempo a usted y a sus anotadores detectar errores y/o lagunas en su modelo o pautas. Debido a esto, sugerimos tener un conjunto de 10 a 20 archivos que pueda usar como un conjunto de capacitación en el que sus anotadores puedan probar los cambios en las pautas. Puede usar estos archivos en su patrón oro más adelante si lo desea o lo necesita, pero debido a la cantidad de veces que sus anotadores habrán repasado esos archivos en particular, probablemente no querrá incluirlos en sus cálculos de IAA.

Además, si descubre que sus pautas han pasado por muchas revisiones, probablemente desee cambiar los archivos de anotaciones de capacitación. De esa forma, sus anotadores no se familiarizarán tanto con las anotaciones esperadas como para dejar de pensar en las pautas. Como regla



general, cambia tus archivos de entrenamiento cada dos o tres veces que revises tus pautas.

## Escribir las pautas de anotación

Las pautas de anotación son las instrucciones sobre cómo desea que sus anotadores apliquen el modelo (su esquema) a los datos (su corpus) para una tarea específica. Las opiniones varían sobre cuánta información debe incluirse en las pautas de anotación: [Dipper et al. \(2004\)](#), por ejemplo, creen que las pautas de anotación deben incluir una descripción del corpus, las teorías lingüísticas que subyacen al modelo, etc. Sin embargo, recomendamos que las pautas de anotación contengan solo la información necesaria para que los anotadores apliquen los conjuntos de etiquetas para anotar el corpus para su tarea de anotación particular (recuerde: la misma especificación se puede usar para diferentes objetivos).

---

### ADVERTENCIA

¡El hecho de que digamos que las pautas de anotación son solo para crear anotaciones no significa que no sea importante realizar un seguimiento de otra información y hacerla disponible más adelante! Es solo que cuando se trata de crear un corpus anotado, incluir demasiada información adicional puede dificultar que los anotadores descubran qué partes del documento son relevantes para su tarea. La información sobre el origen del corpus y cómo se seleccionó, qué teorías lingüísticas se usaron para crear su conjunto de etiquetas y cómo se puede leer y acceder al corpus debe ser absolutamente monitoreada y puesta a disposición como parte del paquete completo del corpus anotado. No lo incluya en las pautas a menos que sea directamente relevante para la forma en que desea anotar el texto. [Capítulo 9](#).

---

Entonces, para hacer las pautas más claras que pueda, hay algunos conceptos básicos que deben cubrir sus pautas de anotación. Las directrices deben responder, como mínimo, a las siguientes preguntas:

- ¿Cuál es el objetivo del proyecto?
- ¿Cómo se llama cada etiqueta y cómo se usa? (Sea específico: proporcione ejemplos y discuta las áreas grises).
- ¿Qué partes del texto desea que se anoten y qué debe dejarse en paz?
- ¿Cómo se creará la anotación? (Por ejemplo, explique qué etiquetas o documentos anotar primero, cómo usar las herramientas de anotación, etc.)

Estas viñetas proporcionan un punto de partida para la mayoría de las tareas, y en el resto de esta sección vamos a repasar cómo podrían ser las respuestas a estas preguntas para algunas tareas diferentes, así como a discutir cómo deben ser las preguntas. modificado para diferentes tipos de tareas. Específicamente, vamos a ver las clasificaciones de etiquetas únicas (críticas de películas), clasificaciones de etiquetas múltiples (géne-

ros cinematográficos), anotaciones de extensión (NE) y anotaciones vinculadas (roles semánticos). Antes de comenzar a escribir sus propias pautas, puede resultarle útil consultar las pautas publicadas para las tareas existentes. Consulte [el Apéndice B](#) para ver algunos enlaces a las tareas de anotación existentes.

---

**NOTA**

¡Asegúrate de escribir tus pautas! Los corpus anotados son más útiles cuando las pautas utilizadas para crearlos están completas y disponibles para descargar con el corpus. Tener las pautas a mano hará que sea más fácil para otras personas recrear o ampliar su tarea más adelante si es necesario.

---

### **Ejemplo 1: etiquetas individuales: reseñas de películas**

Un ejemplo clásico de una tarea de clasificación es clasificar reseñas de películas en categorías positivas y negativas para entrenar a un clasificador para determinar si a alguien le gusta o no una película en función de lo que dice sobre ella. AMovie Review Corpus (MRC) ya existe (está disponible para descargar desde

<http://www.cs.cornell.edu/people/pabo/movie-review-data/> y está incluido en el conjunto de datos de Natural Language Toolkit [NLTK]), y los creadores describen la creación del corpus en función de las reseñas de IMDb.com y la selección de solo aquellas reseñas que otorgaron calificaciones numéricas o de estrellas para categorizar automáticamente si las reseñas fueron positivas o negativas ([Pang y Lee 2004](#)). Entonces, el etiquetado inicial para esta tarea en realidad lo hizo, en cierto sentido, el autor de la revisión, y no se requirieron pautas de anotación.

Pero supongamos que desea crear un conjunto de datos más grande y que los humanos creen las etiquetas para las reseñas. Digamos también que su objetivo es entrenar un algoritmo para poder tomar cualquier revisión y etiquetarla como positiva o negativa, como lo hicieron los creadores del MRC. Sin embargo, si vas a IMDb.com en este momento y mire las reseñas de casi cualquier película, encontrará que no todas las reseñas hacen distinciones claras entre lo bueno y lo malo. ¿Qué harías con una reseña como la siguiente?

*Esta película estuvo bien. Los efectos especiales eran buenos, pero la trama no tenía mucho sentido. Los actores eran divertidos, lo que ayudó, pero la música realmente distraía.*

Ahora, esta revisión no incluye una calificación de estrellas, y no es (en nuestra opinión, de todos modos) claramente positiva o negativa. Parece darle a la película un rotundo "meh". Entonces, ¿cómo encaja eso con su objetivo o con los objetivos del MRC? Bueno, en el caso de los investigadores de MRC, solo estaban observando las diferencias entre las críticas positivas y negativas, por lo que recopilaron específicamente solo aquellas

que eran claramente una u otra. Sin embargo, si su objetivo es entrenar un algoritmo para que funcione en cualquier revisión seleccionada al azar, entonces no querrá limitar su corpus de esa manera.

---

#### ADVERTENCIA

¡Asegúrate de que tu corpus coincida con tu objetivo! Discutimos esto en [el Capítulo 2](#), pero ahora es un buen momento para asegurarse de que cualquier cambio que el ciclo MAMA haya realizado en su tarea se refleje en su corpus. De lo contrario, será difícil entrenar un algoritmo de ML más adelante.

---

Entonces, ¿qué se le debe decir a un anotador que haga con la revisión antes mencionada? Una forma de ayudar a tomar esa determinación es observar cómo otras personas manejan el mismo problema. Los sitios web [RottenTomatoes.com](#) y [Metacritic.com](#) tienen editores que asignan etiquetas a reseñas que no proporcionan categorías o calificaciones específicas. Echemos un vistazo a lo que dicen sobre cómo se asignan estas calificaciones:

*... Nuestro personal revisará cada publicación en nuestra lista de Publicaciones de películas [...] en busca de reseñas de Iron Chef vs. Godzilla. Para cada reseña encontrada, tomaremos la puntuación otorgada por el crítico y la convertiremos a una escala de 0 a 100 puntos. (Para aquellos críticos que no proporcionen una puntuación, les asignaremos una puntuación de 0 a 100 en función de la impresión general que dé la reseña). [...]*

— [Ayuda y soporte de Metacritic](#)

Desafortunadamente, eso realmente no brinda pautas sobre cómo calificar una revisión. Veamos RottenTomatoes:

**¿Cómo se determina si una reseña sin estrellas está podrida o fresca?** La mayoría de los críticos de la Sociedad de Críticos de Cine en Línea (OFCS) ingresan sus propias citas y calificaciones. Para los críticos que no ingresan sus propias citas y calificaciones, básicamente depende del juicio de los editores. Toman en cuenta la elección de palabras, la calificación (si corresponde), el tono y quién es el crítico para determinar si una reseña es positiva o no. Si un editor no está seguro acerca de una revisión, se envía a otro editor para una segunda opinión. Las reseñas "delicadas", las reseñas en las que es realmente difícil determinar si el crítico recomienda la película o no, generalmente reciben un Rotten porque si el crítico no tiene la confianza suficiente para dar a la película incluso una recomendación implícita, entonces no deberíamos. cualquiera.

— [Centro de ayuda de RottenTomatoes](#)

Esa es una explicación clara de cómo asignar una etiqueta a una revisión neutral, con una buena razón de cómo se debe tomar la decisión. También encaja con el objetivo declarado del corpus, aunque puede que no encaje con su propia visión de las reseñas de películas. Suponiendo que esté satisfecho con el uso de esa línea de razonamiento, veamos dónde se encuentra el proyecto con respecto a las preguntas planteadas anteriormente.

### ***¿Cuál es el objetivo del proyecto?***

Etiquetar las reseñas de películas como positivas o negativas.

### ***¿Cómo se llama cada etiqueta y cómo se usa?***

Tenemos dos etiquetas, "positiva" y "negativa", y cada reseña se etiquetará con una de ellas, según el tono de la reseña. Las opiniones que no sean específicamente positivas o negativas se etiquetarán como "negativas".

### ***¿Qué partes del texto desea que se anoten y qué debe dejarse en paz?***

A cada revisión se le asignará una única etiqueta, que se aplicará a todo el documento.

### ***¿Cómo se debe crear la anotación?***

En [el Capítulo 5](#) discutimos los diferentes formatos que se pueden usar para capturar este tipo de datos. En esta situación, probablemente usaríamos una hoja de cálculo rellena previamente con los nombres de los archivos que se están examinando, pero esta decisión depende en gran medida de quiénes son sus anotadores y qué funciona mejor para ellos y para usted.

Tenga en cuenta que no vamos a proporcionar una guía completa en este capítulo para ninguna de las tareas, solo una descripción general de alto nivel de los puntos importantes que deben cubrirse para una buena guía de anotación. Como mencionamos antes, consulte [el Apéndice A](#) para obtener una lista de las tareas de anotación existentes y consulte [el Capítulo 10](#) para ver un ejemplo elaborado de una tarea de anotación.

## **Ejemplo 2: etiquetas múltiples: géneros cinematográficos**

Un ejemplo más complicado de clasificación es identificar géneros cinematográficos. Este tipo de clasificación es un excelente ejemplo de una tarea en la que las pautas de anotación pueden cambiar por completo el resultado de la anotación, y tratar de aplicar las pautas puede generar cambios en la especificación.

#### NOTA

Recuerda que dentro del ciclo MATERIA también se encuentra el ciclo MAMA (Modelo-Anotar-Modelar-Anotar). Es muy probable (especialmente si está comenzando su proyecto de anotación desde cero) que necesitará revisar su modelo y las pautas al menos unas cuantas veces antes de poder anotar todo su corpus. Pero no te desanimes; cuanto más refine su anotación, más fácil será definir características para sus algoritmos de ML más adelante.

Si recuerdas de [Clasificación de géneros cinematográficos](#), cuando discutimos las especificaciones para una tarea de clasificación de géneros cinematográficos, usamos la lista de géneros cinematográficos de IMDb, que incluía los siguientes 26 géneros:

Acción	Aventura	Animación	Biografía	Comedia
Delito	Documental	Drama	Familia	Fantasía
cine negro	Programa de juegos	Historia	Horror	Música
Musical	Misterio	Noticias	Telerrealidad	Romance
Ciencia ficción	Deporte	Programa de entrevistas	Suspense	Guerra
occidental				

Esta ciertamente parece una lista razonable, y proviene de uno de los sitios web de referencia de películas más populares (si no el más) en Internet, por lo que, al menos, es un buen punto de partida para su especificación. Así que echemos un vistazo a la lista de preguntas que deben responderse para las directrices.

#### ***¿Cuál es el objetivo del proyecto?***

Etiquetar resúmenes de películas con notaciones de género.

#### ***¿Cómo se llama cada etiqueta y cómo se usa?***

Tenemos 26 etiquetas que se pueden aplicar a cada resumen según sea necesario.

#### ***¿Qué partes del texto desea que se anoten y qué debe dejarse en paz?***

Cada etiqueta se aplicará a todo el documento.

#### ***¿Cómo se debe crear la anotación?***

El software de anotación es probablemente la mejor manera de aplicar varias etiquetas a un documento.

Bueno, ¡eso fue fácil! Excepto... la respuesta a la segunda pregunta, en particular la parte de "cómo se usa", no está bien especificada. Al etiquetar críticas de películas como positivas o negativas, probablemente sea suficiente decir (al menos como punto de partida) que la etiqueta se basará en el tono, y las críticas neutrales se etiquetarán como "negativas". Sin embargo, las etiquetas de género no se excluyen mutuamente, por lo que los anotadores necesitarán pautas más claras sobre cómo y cuándo aplicar cada una. Una pregunta básica que debe responderse es: "¿Existe un número máximo de etiquetas que se pueden aplicar a un documento?" La respuesta a esta pregunta por sí sola puede diferenciar por completo una tarea de anotación de otra, incluso si cada una usa la misma especificación; directrices que especifican un máximo de, digamos, es probable que dos etiquetas por documento devuelvan un corpus muy diferente al de las pautas que no tienen ese límite. Para la tarea imaginaria que describimos aquí, nuestras pautas no especificarán un límite para la cantidad de etiquetas.

Sin embargo, aunque saber cuántas etiquetas se pueden aplicar responde en parte a la pregunta de "cómo" se usan las etiquetas, no aborda todos los aspectos de "cómo" se usan las etiquetas. Otro aspecto que debe tenerse en cuenta es *cuándo* se utilizará cada etiqueta. En la tarea de revisión positiva/negativa, a cada documento se le asignaba una sola etiqueta, y si un documento no era positivo, era negativo: una dicotomía bastante sencilla. Dado que no hay límite para la cantidad de etiquetas de género (o incluso si hubiera un límite), los anotadores necesitarán alguna aclaración sobre cuándo aplicar qué etiquetas.

Al principio, la cuestión de cuándo usar cada etiqueta parece sencilla. Pero considera las dos primeras etiquetas de la lista anterior: acción y aventura. ¿Cuál, sus anotadores querrán saber, es la diferencia? Un rápidoLa búsqueda de Google muestra que esta no es la primera vez que se hace esta pregunta, y el consenso general parece ser que las películas de acción tienden a ser más violentas, mientras que las películas de aventuras generalmente requieren que el protagonista realice algún tipo de viaje a un lugar o situación que no ha tratado previamente. Sin embargo, lo mismoLa búsqueda de Google también revela que otras listas de géneros (como laNetflix usa) simplemente tienen una etiqueta de género llamada "Acción-Aventura", presumiblemente debido a un alto nivel de superposición entre las dos etiquetas.

Supongamos que está satisfecho con la distinción antes mencionada entre acción y aventura, y las incluye en sus pautas como definiciones para que sus anotadores puedan consultarlas. Así que ahora (con suerte) etiquetarán *Die Hard* como una película de acción y *La vuelta al mundo en 80 días* como una aventura. Pero espera un minuto, uno de los primeros resúmenes de *Die Hard* en IMDb.com comienza con "El detective de la ciudad de

Nueva York John McClane acaba de llegar a Los Ángeles para pasar la Navidad con su esposa. Desafortunadamente...." Entonces, ¿hay un viaje involuacrado? El personaje va a un lugar diferente, pero un policía que está acostumbrado a tratar con delincuentes que va a una ciudad diferente para tratar con delincuentes realmente no cumple con la cláusula de "nueva situación" de la definición de aventura dada anteriormente, por lo que probablemente podamos Decir con seguridad que *Die Hard* no califica como una película de aventuras.

Bien, ese es un buen comienzo. Pero, ¿qué pasa con *Eat, Pray, Love* ? El personaje principal claramente emprende un viaje a nuevos lugares y situaciones, pero sospechamos que la mayoría de la gente no la consideraría una película de aventuras. Entonces, ¿quizás las películas de aventuras también tienen que tener algún elemento de peligro? Es mejor modificar la definición en sus directrices. O tal vez en este punto sientes que tratar de diferenciar entre los dos es un poco tedioso y/o sin sentido, y prefieres modificar tu especificación para tener una sola categoría de acción y aventura.

---

#### NOTA

Lo crea o no, no incluimos los párrafos anteriores simplemente porque nos gusta criticar los géneros cinematográficos. Más bien, las incluimos porque esta discusión ilustra los tipos de preguntas que deberá responder para sus anotadores cuando les dé las pautas para su tarea. El enfoque más simple para la tarea es simplemente dar a los anotadores una pila de textos y decirles que coloquen las etiquetas que les parezcan correctas, pero no olvide que una parte importante de una tarea de anotación es *reproducibilidad* \_ Si simplemente les dice a los anotadores que etiqueten lo que quieren, es poco probable que un conjunto diferente de anotadores pueda brindarle los mismos (o incluso similares) resultados en una fecha posterior.

---

Si no está completamente seguro de qué definición dar a cada etiqueta, ahora sería un buen momento para seguir otro consejo que hemos repetido varias veces (y seguiremos repitiendo): ¡investiga un poco! Un excelente libro que encontramos sobre el tema es el de Barry Keith Grant. *Género cinematográfico: de la iconografía a la ideología* (Wallflower Press, [2007](#) ). Si bien no toda la teoría del libro se puede aplicar necesariamente a una tarea de anotación, mirar los diferentes géneros en términos de temas en lugar de simplemente mirar los detalles superficiales puede ayudar a aclarar qué hace que una película encaje en un género. Por ejemplo, las películas del oeste a menudo tratan, en parte, sobre la exploración de nuevas fronteras y el espíritu pionero, una definición que podría ser más efectiva y relevante que una que especifica que un western tiene caballos, personas con sombreros de vaquero y al menos una persona que dice “pardner” muy lentamente.



Una mirada más cercana a los géneros cinematográficos también revela que no todos los géneros de la lista describen los mismos aspectos de la película. Mientras que las etiquetas como "Acción", "Aventura", "Crimen" y "Romance" le dicen al lector algo sobre los eventos que tendrán lugar en la película, las etiquetas "Histórico", "Ciencia ficción" y "Fantasía". se refieren al escenario, y "Animación", "Talk-Show" y "Reality-TV" describen las circunstancias de producción. Por lo tanto, un enfoque completamente diferente para esta tarea sería dividir estos géneros en categorías (producción, escenario, etc.) y pedirles a los anotadores que asignen al menos una etiqueta de cada categoría. Suponiendo que las categorías y etiquetas estén suficientemente bien definidas (lo que no es necesariamente una tarea fácil), el requisito específico de "al menosX número de etiquetas" puede mejorar en gran medida el IAA y la reproducibilidad de su tarea. Si tuviera que adoptar este enfoque, podría crear un DTD (Definición de tipo de documento) que se parece a esto:

```
<!Configuración de ELEMENTO (VACÍO) >
<!ATTLIST descripción de la configuración (histórica | ciencia ficción | fantasía) >

<!ELEMENTO producción (VACÍO) >
<!ATTLIST circunstancias de producción (animación | documental |
    programa de juegos | musicales | noticias | telerrealidad | programa de entrevistas) >

<!ELEMENTO contenido (VACÍO) >
<!ATTLIST tipo de contenido ( acción | aventura | biografía |
    crimen | teatro | misterio | romanticismo) >
```

Por supuesto, esta DTD reorganizada es solo un comienzo: si desea exigir que a cada película se le asigne una configuración, necesitará al menos una que pueda describir una película ambientada en el presente sin aumentos particulares a la realidad. Pero esta es otra forma de enmarcar la tarea de género que podría resultar más útil, según el objetivo que te hayas fijado.

En general, es importante darse cuenta de que si tiene una tarea con muchas etiquetas diferentes que desea usar, es vital que cree definiciones claras para cada etiqueta y brinde ejemplos de cuándo desea que sus anotadores usen cada una de ellas (e igualmente importante, cuando *no* desea que sus anotadores los usen). Si bien esto puede parecer más importante cuando crea sus propias etiquetas en lugar de basarse en los términos existentes, también desea asegurarse de que los juicios de los anotadores no se vean empañados por sus propias ideas preconcebidas sobre el significado de un término.

Otra posible causa de confusión para los anotadores, además de su conocimiento sobre un término en la especificación, es su conocimiento del material que se está anotando. Si le está pidiendo a un anotador que cree etiquetas que describan una película basándose en un resumen escrito, pero el anotador ha visto la película y siente que el resumen es inexacto, deberá abordar si puede usar su propio conocimiento del mundo para aumentar la información en el documento que está etiquetando. Aunque para fines de reproducibilidad y ML, recomendamos encarecidamente no utilizar el conocimiento o la intuición externos como fuente de datos de anotación.

---

### Ejemplo 3: Anotaciones de extensión: entidades con nombre

En **Adición de entidades con nombre**, proporcionamos una especificación de muestra para NE relacionados con películas: título de la película, director, escritor, actor y personaje. Si bien las definiciones de cada uno de estos roles deberían ser un poco más fáciles de definir que los géneros discutidos en la sección anterior, las etiquetas NE son etiquetas de *extensión*, lo que significa que se aplican a tramos de texto específicos dentro de un documento, en lugar de al documento como un entero. Las etiquetas de extensión traen consigo un nuevo conjunto de consideraciones, como las siguientes:

- ¿Cómo deben decidir los anotadores cuánto debe durar cada intervalo etiquetado? ¿Cuántas palabras se incluyen en cada etiqueta? Por ejemplo, ¿deberían incluirse frases descriptivas? ¿Qué pasa con los títulos, honoríficos o determinantes? “Una casa blanca” es claramente diferente de “la Casa Blanca”, pero ¿es el determinante realmente la parte importante de esa frase?
- ¿Qué pasa con los nombres que se dividen? Si bien es razonable etiquetar a “Los Wachowski” como una sola entidad, ¿qué pasa si el texto dice “Lana y Andy Wachowski”? ¿Se etiquetará toda la frase o desea que sus anotadores hagan algo diferente en ese caso?
- ¿Sus anotadores deberían anotar cada mención de un NE, o solo la primera vez que se menciona?
- ¿Qué pasa si una entidad parece estar desempeñando dos roles diferentes? En una anotación NE estándar, es posible que vea “Ayuntamiento de Boston”; ¿debería marcarse “Boston” en la misma medida que “Ayuntamiento” o son diferentes? ¿Se superpone uno al otro?
- ¿Qué pasa con las construcciones posesivas, como “*The Breakfast Club* de John Hughes”? Y si decide que “John Hughes” debe etiquetarse por separado de “*The Breakfast Club*”, ¿la anotación incluye el apóstrofe?

Las preguntas anteriores deberán abordarse para prácticamente cualquier tarea de NE y, en mayor medida, para cualquier tarea que involucre anotaciones de extensión, y la mayoría de ellas se reducen a esto: ¿Cuál es *el alcance de cada etiqueta*? Es decir, ¿cuáles son las limitaciones y pautas sobre dónde se debe colocar cada etiqueta y cuántas veces se debe usar?

Afortunadamente, los NE han sido parte de las anotaciones de lenguaje natural durante años, por lo que si está dispuesto a investigar un poco, encontrará las respuestas a estas preguntas que otros investigadores han encontrado útiles. Una tarea NE notable tuvo lugar en el Message Understanding Conference (MUC) 6 en 1995. Las pautas para la anotación NE que se formaron allí se han modificado para varias otras anotaciones a lo largo de los años, y el El Consorcio de datos lingüísticos (LDC) actualmente tiene algunas buenas pautas para realizar la anotación NE, que puede encontrar en <http://projects.ldc.upenn.edu/LCTL/Specifications/> . Si bien estas pautas se enumeran para idiomas distintos del inglés, las pautas mismas están en inglés y brindan ejemplos en inglés.

Notarás que el El documento Simple Named Entity Guidelines V6.5 de LDC contiene una sección llamada "Casos difíciles", que aborda las preguntas sobre la anotación NE que planteamos anteriormente. Esta es una excelente manera de estructurar sus pautas: hacer que la primera parte brinde ejemplos claros de cómo se usa una etiqueta y luego mantener los ejemplos complicados en un solo lugar. Esto hará que sea mucho más fácil para sus anotadores encontrar lo que necesitan rápidamente mientras crean sus anotaciones, y será mucho menos frustrante para ellos.

Por supuesto, muchos de los puntos que mencionamos sobre las anotaciones en los otros ejemplos de pautas también se aplican (¿Qué pasa si alguien es tanto director como escritor? ¿Obtiene dos etiquetas? Si está realizando una tarea de NE más general, ¿son cada uno de los diferentes tipos de entidades claramente definidos para que su anotador sepa cuándo usar cada uno), pero no vamos a repasarlos todos nuevamente aquí.

Cuando se trata de etiquetas de extensión, es importante que las pautas expliquen el alcance de la etiqueta: dónde se debe (y no se debe) aplicar la etiqueta y qué hacer cuando aparece un fragmento de texto relevante en diferentes contextos. Esto es particularmente importante porque, dependiendo de cómo calcule el IAA, las diferencias en el lugar donde comienzan y terminan las anotaciones pueden tener efectos masivos en las puntuaciones del acuerdo.

#### **Ejemplo 4: etiquetas de enlace: funciones semánticas**

Si bien la adición de etiquetas de extensión a una tarea de anotación requiere pautas claras sobre dónde deben comenzar y terminar esas etiquetas, las etiquetas de enlace traen dos nuevas preguntas a las pautas que deben responderse:

- ¿Cuáles son los enlaces que conectan?
- ¿Cuándo se debe crear un enlace?

Estas preguntas pueden parecer bastante sencillas, pero recuerde nuestro ejemplo de una anotación de enlace temporal de la discusión sobre informatividad y corrección en [Refinar su objetivo: Informatividad versus corrección](#) . Es cierto que ese fue un ejemplo un tanto extremo de cómo

una tarea de anotación puede salirse de control, pero ilustra la importancia de tener pautas claras sobre cuándo se necesitan enlaces y cuándo se deben crear. Afortunadamente, la mayoría de las tareas de anotación tienen límites mucho más claros cuando se trata de vincular extensiones que nuestra anotación temporal de muestra. Por ejemplo, la tarea de rol semántico para películas que discutimos en **Roles semánticos** no tiene tanto potencial para volverse completamente fuera de proporción, aunque también tiene potencial para la confusión. Recuerde que la tarea especificaba que las relaciones semánticas entre actores, personajes, escritores, directores y películas se anotarían con los roles `acts_in`, `acts_as`, `directs`, `writes` y `character_in`. Entonces, por ejemplo, en la oración "James Cameron dirigió *Avatar*" tendríamos un enlace entre "James Cameron" (que sería etiquetado como `director`) y "*Avatar*" (que sería etiquetado como `film_title`), y el enlace sería `directs`.

Pero incluso este ejemplo sencillo tiene algunos lugares donde la tarea podría volverse más complicada. Veamos la revisión de muestra que vimos en el capítulo anterior (aquí los actores, escritores y directores están en **negrita**, los títulos de las películas están en *cursiva* y los personajes están en `constant width`):

*En Love, Actually, el escritor y director **Richard Curtis** teje una historia enrevesada sobre los personajes y sus relaciones. De particular interés es **Liam Neeson** ( *La lista de Schindler*, *Star Wars* ) como *Daniel* un hombre que lucha por lidiar con la muerte de él *wife* y la relación con su joven hijastro *Sam* ( **Thomas Sangster** ). **Emma Thompson** ( *Sense and Sensibility*, *Henry V* ) brilla como una *mujer middle-aged housewife* cuyo matrimonio con ella *husband* (interpretado por **Alan Rickman** ) está bajo el asedio de un *beautiful secretary*. Si bien esta película tiene sus momentos puramente cómicos (principalmente presentados por **Bill Nighy** como una estrella de rock obsoleta *Billy Mack*), esta película evita la comedia más directa que **Curtis** ha presentado antes como escritor de *Blackadder* y *Mr. Bean*, presentando en cambio una visión notable y suavemente humorística de lo que es el amor en realidad.*

Si bien la mayoría de las anotaciones de roles semánticos aquí son bastante sencillas, hay algunas piezas que pueden hacer tropezar a los anotadores concienzudos. Por ejemplo, al crear `writes` enlaces para *Blackadder* y *Mr. Bean*, ¿deberían vincularse esos títulos de películas al **Curtis** que aparece en la misma oración, o deberían vincularse nuevamente al **Richard Curtis** en la primera oración, porque ese es su nombre completo? Del mismo modo, ¿deberían vincularse todas `act_in` y cada una `character_in` de las relaciones de la película que se está revisando con la mención del título en la primera oración, o deberían vincularse con frases (actualmente sin anotaciones) como "esta película"? Si *Amor, en reali-*

*dad* se mencionaron más de una vez en la reseña, ¿los anotadores deberían vincular actores y personajes a la mención más cercana del título, solo a la primera, o a todos?

No le daremos las respuestas a estas preguntas, porque no hay una respuesta verdadera para ellas. La forma en que aborde la anotación dependerá de su objetivo y modelo, y la simple prueba y error con sus pautas y anotadores ayudará a determinar cuál es la respuesta más razonable y útil a estas preguntas para su tarea. Sin embargo, no olvide consultar las pautas para tareas similares para obtener sugerencias sobre lo que ha funcionado para otras personas.

## **Anotadores**

Un componente clave de cualquier proyecto de anotación es, por supuesto, las personas que realizan la tarea de anotación. Claramente, esto significa que se debe pensar en quién encuentra para crear sus anotaciones. Para ello, sugerimos poder responder a las siguientes preguntas:

- ¿Qué idioma o idiomas necesitan saber sus anotadores para realizar su tarea de anotación?
- ¿Su tarea de anotación requiere algún conocimiento especializado para comprender o realizar?
- ¿Cuáles son las consideraciones prácticas que deben tenerse en cuenta (dinero, tiempo, tamaño del conjunto de datos, etc.)?

Repasemos estos uno a la vez.

### ***¿Qué idioma o idiomas necesitan saber sus anotadores para realizar su tarea de anotación? Y además, ¿qué tan bien necesitan conocerlos?***

Lo más probable es que la respuesta a esta pregunta sea bastante obvia, pero vale la pena especificarla. Si su tarea requiere una lectura atenta de un texto (p. ej., relaciones anafóricas, desambiguación del sentido de las palabras o roles semánticos), es posible que desee limitar sus anotadores a hablantes nativos del idioma que está anotando. Sin embargo, para algunas anotaciones, es posible que pueda utilizar hablantes no nativos, y para algunas tareas incluso podrían preferirse (p. ej., si el propósito de la tarea es conocer las percepciones del estudiante de un segundo idioma sobre su nuevo idioma). Independientemente de lo que decida, asegúrese de dejar claras las preferencias de idioma en cualquier publicación o descripción de trabajo.

### ***¿Su tarea de anotación requiere algún conocimiento especializado para comprender o realizar?***

Además de los idiomas en los que se encuentran los textos, ¿hay algún otro conocimiento externo que sus anotadores deban tener para realizar bien esta tarea? Si su tarea es el etiquetado de POS,

encontrar anotadores que estén familiarizados con esos conceptos (quizás personas que hayan tomado uno o dos cursos de sintaxis) probablemente reducirá el tiempo necesario para capacitar a sus anotadores y aumentará el IAA.

Hay otros factores que pueden afectar lo que sus anotadores necesitan saber para desempeñarse bien en su tarea de anotación, como el material de origen real. Las anotaciones biomédicas y clínicas son áreas en las que cada vez más investigadores de procesamiento del lenguaje natural (NLP) están investigando, pero es mucho más fácil para un anotador identificar y etiquetar expresiones génicas en artículos científicos si ya está familiarizado con los conceptos y el vocabulario. Los documentos clínicos, como las notas del hospital y los resúmenes de alta, pueden ser aún más complicados, porque es probable que necesite a alguien capacitado como RN (si no como MD) para interpretar cualquier información médica que pueda interesarle debido a la densidad y la jerga. el texto es.

---

#### NOTA

Si decide que seleccionará anotadores con determinadas habilidades o conocimientos, asegúrese de realizar un seguimiento de esa información y ponerla a disposición de otras personas que utilicen su corpus y sus directrices. La reproducibilidad de una tarea de anotación aumenta cuando se tienen en cuenta todas las variables, ¡como cualquier otro experimento!

---

### ***¿Cuáles son las consideraciones prácticas que deben tenerse en cuenta?***

Una cosa que debe tener en cuenta al planificar su proyecto de anotación y dónde encontrar anotadores es que la anotación lleva tiempo. Obviamente, las tareas que tienen una alta densidad de etiquetas, como el etiquetado de POS, consumen mucho tiempo simplemente porque hay una proporción uno a uno de etiquetas a palabras. Pero más que eso, la mayoría de las personas solo pueden realizar la mayoría de las tareas de anotación durante unas pocas horas a la vez. La anotación requiere mucha concentración y atención a los detalles, y si espera que sus anotadores lo hagan de 9:00 a 5:00 durante días seguidos, es probable que obtenga anotaciones muy inconsistentes. La anotación se acelerará a medida que sus trabajadores se acostumbren a la tarea, pero asegúrese de dejar suficiente tiempo en su agenda para que sus anotadores hagan un buen trabajo.

---

#### NOTA

Si está ampliando una tarea/guía de anotación que ya existe, vale la pena entrenar a sus anotadores en los datos del conjunto de datos anterior. De esa manera, tiene una forma sólida de evaluar si sus anotadores entienden la tarea dada y puede hacer los ajustes necesarios a las pautas sin comprometer su propio conjunto de datos.

---

En teoría, si tuviera una agenda apretada, simplemente podría contratar y capacitar a más anotadores para que trabajen todos al mismo tiempo. Sin embargo, como veremos más adelante en **Evaluación de las anotaciones**, debe asegurarse de que cada archivo se anote al menos dos veces (para que pueda calcular las puntuaciones de IAA), y estas cosas generalmente son más fáciles de administrar cuando no está abrumado con los anotadores. .

Además, incluso si sus pautas de anotación se han modificado y perfeccionado repetidamente, cuanto más tiempo tenga un anotador para adaptarse a una tarea, mejor lo hará, y cuanto más tiempo dedique a realizar la anotación, mejores serán sus anotadores. podrá aclimatarse a la tarea y, por lo tanto, generar anotaciones más precisas.

---

### DÓNDE BUSCAR ANOTADORES

Suponiendo que no tenga compañeros de trabajo que use para anotar sus textos, necesitará contratar anotadores en algún momento. Si bien ciertamente es posible publicar un anuncio en Craigslist para anotadores, para tareas de anotación que no requieren ningún entrenamiento especial más que conocer el idioma del texto, recomendamos verificar primero para ver si hay colegios o universidades locales que tengan conocimientos lingüísticos. o programas de lingüística computacional. Según la tarea, es posible que ni siquiera sea necesario exigir que un anotador sea un lingüista, pero si su tarea involucra algo que tiene cierta complejidad lingüística (como el etiquetado de POS), es útil buscar anotadores que ya serán familiarizado con los conceptos básicos.

Por otro lado, para tareas que se basan principalmente en la simple interpretación del texto (p. ej., determinar si la crítica de una película es positiva o negativa), probablemente no sea necesaria una formación en lingüística.

Finalmente, si su tarea requiere algún tipo de capacitación especial para interpretar (tal vez esté interesado en textos de un campo específico, como la física o la bioestadística), entonces seguramente se beneficiará al contratar anotadores que tengan la capacitación necesaria para comprender su tarea. Una vez más, los colegios o universidades locales suelen ser un buen lugar para comenzar.

La Lista LINGUIST mantiene una base de datos de lingüística y programas relacionados con la lingüística en todo el mundo. Puede acceder a él en <http://linguistlist.org/teach/programs/>.

---



# Elección de un entorno de anotación

Muchos entornos de anotación diferentes están disponibles para su uso, desde herramientas independientes y gratuitas, como elEntorno de anotación multipropósito (MAE) yCallisto, a bancos de trabajo que brindan administración de tareas de anotación y opciones de adjudicación, todo empaquetado (como PIZARRA y laHerramienta de anotación de Brandeis [BAT]). Se proporciona una guía de usuario para MAE en [el Apéndice C](#) , y MAE es una buena herramienta para comenzar si nunca antes ha realizado anotaciones, pero los proyectos de anotación no son una herramienta única para todos. No vamos a revisar la lista completa de herramientas aquí (una lista en el [Apéndice B](#) brinda una descripción general de lo que está disponible y algunas de las características básicas de cada pieza de software), pero hay algunas preguntas que debería poder responder. para responder sobre su tarea, y eso ayudará a determinar qué software de anotación será mejor para usted.

---

## ADVERTENCIA

No olvide que también deberá dar a sus anotadores instrucciones sobre cómo desea que usen la herramienta de anotación que elija para la tarea de anotación, y probablemente también proporcione al menos un tutorial en persona. Elegir una herramienta de anotación con una curva de aprendizaje muy pronunciada para los anotadores, o una en la que es fácil cometer errores, también puede causar errores en la anotación y puntuaciones de concordancia resultantes más bajas.

---

Siempre hay consideraciones básicas en las que debe pensar cuando se trata de elegir un software. Por ejemplo, ¿está siendo compatible actualmente, por lo que si tiene preguntas al respecto, habrá una manera de encontrar una respuesta? ¿Funcionará en todas las computadoras y sistemas operativos que usted y sus anotadores usarán? Si el software es de código abierto y desea realizar un cambio, ¿tiene a alguien que pueda cambiar el código para realizar los ajustes que desea? (Esto no es necesario para usar software de código abierto, por supuesto, pero puede ser una buena manera de obtener la anotación que desea). Sin embargo, estas preguntas son bastante genéricas, por lo que aquí hay algunas preguntas específicas de anotación. eso afectará su elección de software.

### ***¿Qué partes de su tarea necesita respaldar el software de anotación?***

Si está anotando extensiones en un texto y desea que sus anotadores puedan modificar fácilmente los atributos asociados con las etiquetas que crean, querrá asegurarse de que la herramienta que elija haga que esa funcionalidad sea de fácil acceso. No todo el software de anotación (o incluso la mayoría) se creó para ser completamente multipropósito: algunos software no permiten a los usuarios



crear enlaces entre extensiones anotadas, y si desea poder dar a sus anotadores acceso a ontologías u otros recursos existentes mientras funcionan, tendrá que elegir un software que lo tenga incorporado.

### ***¿Cuáles son las unidades de su tarea de anotación?***

En **Anotación de extensión de texto: entidades con nombre**, discutimos las diferentes formas en que se puede representar la anotación de extensión, como basada en tokens o basada en caracteres. Los estándares actuales favorecen las anotaciones separadas basadas en caracteres, pero no todas las herramientas usan ese sistema para su representación de anotaciones. Esto no es necesariamente un problema si no está comprometido con el paradigma de separación o si puede convertir de un formato a otro más adelante. Sin embargo, como discutimos antes, algunas tareas requieren que se anoten palabras parciales y herramientas que solo anotan a nivel de token, por lo que si tiene una de esas tareas, querrá evitar el software con esa restricción.

### ***¿Desea que sus anotadores realicen todas las partes de la tarea de anotación a la vez o debe dividir la tarea en capas?***

Algunas herramientas de anotación (como BAT) imponen un proceso de anotación en capas, lo que significa que cada tipo de etiqueta se anota por separado, con etiquetas de extensión que se anotan y adjudican primero, y luego vinculan las etiquetas creadas sobre las extensiones adjudicadas. Este formato significa que los enlaces tendrán inherentemente puntajes de adjudicación más altos, porque los anotadores trabajarán desde el mismo conjunto de extensiones anotadas. Sin embargo, si todavía está resolviendo los problemas en su tarea de anotación, este podría no ser el mejor formato para usar, porque requiere que se realicen grandes porciones de la anotación y la adjudicación antes de pasar al resto de la anotación. lo que significa que si hay un error en sus pautas relacionadas con las etiquetas de extensión que afectan la forma en que se crean los enlaces, tendrá que volver a hacer mucho trabajo para corregir el error. Sin embargo, la anotación en capas puede ser una excelente manera de dividir una tarea en partes manejables, y el paradigma se puede usar incluso si no está usando una herramienta que aplique ese formato.

### ***¿Qué otras características quieres poder incorporar?***

Como mencionamos al comienzo de esta sección, algunas herramientas brindan funciones de administración de tareas (para administrar qué anotadores obtienen qué archivos, realizar un seguimiento de la precisión, etc.), mientras que otras brindan interfaces de anotación simples. Sin embargo, algunas herramientas (como GATE) también integran etiquetadores POS, tokenizadores y algoritmos ML. Estas funciones ciertamente no son obligatorias para las

tareas de anotación, pero pueden proporcionar algunas herramientas útiles para los anotadores (y para usted, por supuesto).

## Evaluación de las anotaciones

Antes de comenzar a crear el estándar de oro adjudicando las anotaciones, querrá evaluar la tarea de anotación, especialmente si todavía está trabajando en el ciclo MAMA con su conjunto de prueba de archivos de anotación. Hemos hablado mucho hasta ahora sobre Puntuaciones IAA (a veces denominadas puntuaciones de acuerdo *entre codificadores* o *entre etiquetadores*) y cómo es importante tener buenas puntuaciones IAA para asegurarse de que su tarea sea reproducible. Las puntuaciones IAA proporcionan una forma de evaluar la precisión con la que dos o más anotadores pueden realizar su tarea de anotación.

---

### ADVERTENCIA

Las buenas puntuaciones de IAA no significan necesariamente que su corpus producirá buenos resultados cuando se use para entrenar un algoritmo de ML. Sin embargo, cuantos más datos pueda anotar, mejores serán sus resultados de ML, por lo que aún vale la pena dedicar tiempo a hacer que una tarea sea reproducible: cuanto más fácil sea su tarea para otros, más anotadores podrá entrenar, para que pueda tener un corpus más grande para pruebas y entrenamiento.

---

Al principio, puede parecer que calcular el IAA es solo una cuestión de contar cuántas etiquetas hay en el conjunto de datos y calcular cuántas veces los anotadores están de acuerdo sobre si cada etiqueta debe estar allí. Sin embargo, el uso de porcentajes directos como ese no tiene en cuenta los acuerdos aleatorios que probablemente ocurran cuando las personas anotan textos. Si, por ejemplo, desea que un conjunto de reseñas de películas se etiqueten como positivas o negativas, si sus anotadores simplemente eligen la etiqueta para cada documento sin leer el texto, hay un 50 % de posibilidades de que estén de acuerdo con la etiqueta que asignan. lo que significa que las puntuaciones de concordancia parecerán artificialmente altas.

Para crear números que puedan compararse entre estudios, se han desarrollado y utilizado varias métricas a lo largo de los años para calcular el IAA. Los dos más utilizados en lingüística computacional y de corpus son *el Kappa de Cohen* (y su variación, *el Kappa de Fleiss*) y *el Alpha de Krippendorff*. Todavía se está debatiendo si estas métricas miden adecuadamente o no todos los aspectos del acuerdo en una tarea de anotación (para obtener excelentes revisiones del tema, consulte [Artstein y Poesio 2008](#) y [Bayerl y Paul 2011](#)), pero aquí solo cubriremos los conceptos básicos de la métrica kappa, ya que se aplican a la mayoría de las tareas de anotación.

## Kappa de Cohen ( $\kappa$ )

El Kappa de Cohen ( $\kappa$ ) mide la concordancia entre dos anotadores, teniendo en cuenta la posibilidad de concordancia fortuita. la ecuación es:

$$K = \frac{\text{Pr}(a) - \text{Pr}(e)}{1 - \text{Pr}(e)}$$

En la ecuación,  $\text{Pr}(a)$  es la concordancia relativa observada entre los anotadores, y  $\text{Pr}(e)$  es la concordancia esperada entre los anotadores, si cada anotador eligiera aleatoriamente una categoría para cada anotación.

Echemos un vistazo a un ejemplo específico.

Volviendo a uno de los ejemplos que hemos estado discutiendo, supongamos que teníamos una tarea de anotación en la que se pedía a dos anotadores, A y B, que asignaran las etiquetas "positivo", "neutro" y "negativo" a un conjunto de 250 reseñas de películas, y que las anotaciones resultantes se veían así:

		B	B	B
		positivo	neutral	negativo
A	positivo	54	28	3
A	neutral	31	18	23
A	negativo	0	21	72

La tabla anterior muestra que, si bien los comentaristas rara vez tuvieron situaciones en las que uno etiquetó una revisión como "positiva" y el otro como "negativa", la etiqueta "neutral" contribuyó mucho al nivel de desacuerdo. Pero, ¿cómo convertimos estos números en una métrica de evaluación?

---

### NOTA

Esta tabla es esencialmente una *matriz de confusión*, que es una tabla que se usa para comparar la salida de un algoritmo con un estándar de oro. Hablaremos más sobre las matrices de confusión en el [Capítulo 8](#).

---

Primero, calculamos  $\text{Pr}(a)$ , el acuerdo real entre los anotadores. De 250 documentos, A y B dijeron "positivo" 54 veces, ambos dijeron "neutral" 18 veces y ambos dijeron "negativo" 72 veces. Entonces el porcentaje de acuerdo observado es:

$$\text{Pr}(a) = (54 + 18 + 72) / 250 = .576 \text{ (57.6\%)}$$

A continuación, calculamos  $Pr(e)$ , la concordancia de probabilidad esperada, para cada etiqueta. Para hacer eso, determinamos el porcentaje de tiempo que cada anotador usó cada etiqueta y multiplicamos esos porcentajes para determinar con qué frecuencia los dos anotadores usarían la misma etiqueta en el mismo documento al mismo tiempo, luego sumamos cada uno de ellos para obtener  $Pr(e)$ . Si eso suena confuso, solo mire:

A utilizó la etiqueta "positivo" 85 veces ( $54 + 28 + 3$ ), o el 0,34 % de las veces. B también usó la etiqueta "positivo" 85 veces ( $54 + 31$ ), que también es .34. Multiplicados,  $.34 \times .34 = .116$ , por lo que A y B tienen una probabilidad de **.116** de elegir aleatoriamente "positivo" como etiqueta.

Ahora hacemos los mismos cálculos para "neutro" y "negativo". A usó "neutral" 72 veces, o 0,288 %, y B usó la etiqueta 67 veces, o 0,268 %. Combinados,  $(.268 \times .288)$ , hay una probabilidad de **.077** de que ambos usen la etiqueta "neutral". Finalmente, A usó "negativo" 93 veces (.372) y B usó "negativo" 98 veces (.392), dando una probabilidad de acuerdo de **.146** en la etiqueta "negativo".

La suma de esos tres puntajes de acuerdo de probabilidad juntos nos da:

$$Pr(e) = .116 + .077 + .146 = \mathbf{.339}$$

Poner  $Pr(a)$  y  $Pr(e)$  en la ecuación nos da:

$$\kappa = (.576 - .339) / (1 - .339) = .237 / .661 = \mathbf{.359}$$

Discutiremos cómo interpretar las puntuaciones en **Interpretación de los coeficientes Kappa**; por ahora, pasemos a otro ejemplo.

### **Kappa de Fleiss ( $\kappa$ )**

El Kappa de Cohen puede determinar el acuerdo entre dos anotadores, pero ¿qué pasa si hay tres o más personas anotando el mismo documento? En ese caso, deberá usar el Kappa de Fleiss en su lugar. El Kappa de Fleiss no se basa en realidad en el Kappa de Cohen; más bien, es una extensión del  $\pi$  ( $\pi$ ) de Scott, pero para la mayoría de los contextos, son lo suficientemente similares como para compararlos.

La ecuación base para el  $\kappa$  de Fleiss es esencialmente la misma que para el  $\kappa$  de Cohen, con el acuerdo real y el acuerdo esperado debido al azar calculados y comparados. Sin embargo, la forma en que se calculan difiere, por lo que usaremos símbolos ligeramente diferentes aquí para evitar confusiones:

$$k = \frac{P - P_e}{1 - P_e}$$

La tabla utilizada para representar los valores del anotador para el Kappa de Fleiss, en lugar de tener un eje por anotador, tiene un eje para los valores posibles que un anotador podría asignar y el otro eje para cada uno de los elementos que se anotan. El contenido de las celdas muestra cuántos anotadores asignaron cada categoría a cada elemento. Tenga en cuenta que Fleiss's Kappa no asume que todos los elementos están anotados por los *mismos* anotadores, pero sí asume que todos los elementos están anotados la misma cantidad de veces. Solo para que podamos ver algunos números más grandes, supongamos que rehicimos nuestra tarea de anotación de reseñas de películas como un proyecto de crowdsourcing (consulte el [Capítulo 12](#)). En lugar de tener 250 reseñas de películas anotadas por 2 personas, digamos que tenemos 5 reseñas de películas anotadas como positivas, neutrales o negativas por 250 personas cada una. Estas anotaciones se representarían así:

	positivo	neutral	negativo
Opinión 1	85	72	93
Revisión 2	85	67	98
Reseña 3	68	99	83
Reseña 4	88	88	74
Reseña 5	58	120	72
total	384	446	420

En esta tabla, las categorías están en la parte superior, los documentos de reseñas de películas están al costado y el contenido de cada celda representa cuántas veces un anotador asignó cada categoría a cada reseña.

Primero, necesitamos calcular cuántas asignaciones fueron (proporcionalmente) a cada categoría. Esto está representado por  $P_c$ , donde  $c$  representa la categoría que se está evaluando. Hacemos esto sumando el contenido de cada fila y dividiendo por el número total de anotaciones. En la siguiente ecuación,  $A$  es el número de reseñas,  $a$  es el número de anotaciones por reseña,  $k$  es el número de categorías e  $i$  representa la celda actual de la tabla:

$$P_c = \frac{1}{Aa} \sum_{i=1}^A a_{ic}, \quad 1 = \frac{1}{a} \sum_{c=1}^k a_{ic}$$

Esto parece un poco complicado, pero básicamente todo lo que dice es que  $P_c$  (en este caso,  $c$  representará positivo, neutral y negativo) será igual a la suma de los valores en su columna dividida por el número de revisiones (5) veces el número de anotaciones por revisión (250). La segunda versión de la ecuación simplemente dice que si sumas todas las anotaciones que hizo un anotador y las divides por el número de anotaciones que cada anotador hizo individualmente, obtendrás 1, porque ambos números deben ser iguales.

Entonces, si aplicamos la ecuación  $P_c$  a la primera categoría de anotación, obtenemos la siguiente ecuación:

$$P(\text{positivo}) = (85 + 85 + 68 + 88 + 58) / (5 \times 250)$$

$$= 384 / 1250$$

$$= .3072$$

Si aplicamos este cálculo al resto de la tabla, podemos llenar la fila inferior de la tabla, así:

	positivo	neutral	negativo
Opinión 1	85	72	93
Revisión 2	85	67	98
Reseña 3	68	99	83
Reseña 4	88	88	74
Reseña 5	58	120	72
total	384	446	420
<b>PC _</b>	<b>.3072</b>	<b>.3568</b>	<b>.336</b>

A continuación, necesitamos calcular  $P_i$ , que representa el acuerdo del anotador por revisión en comparación con todos los valores de acuerdo posibles. Como antes,  $a$  es el número de anotaciones por revisión,  $k$  es el número de categorías,  $c$  es la categoría actual e  $i$  es la revisión actual.

$$P_i = \frac{(\sum_{c=1}^k a_{ic}^2) - (a)}{a(a-1)}$$

Nuevamente, esto no es tan difícil de aplicar como parece: para cada fila, solo estamos sumando los cuadrados de los valores en cada columna y moderando el resultado por el número total de anotaciones para cada revisión. Entonces, para la Revisión 1, calcularíamos esto:

$$P(\text{Revisión 1}) = (85^2 + 72^2 + 93^2) - 250 / 250(250-1)$$

$$= 21058 - 250 / 62250$$

$$= 20808/62250$$

$$= \mathbf{.3343}$$

Al realizar este cálculo para cada fila de Revisión en la tabla, podemos completar la última columna, así:

	positivo	neutral	negativo	Pi
Opinión 1	85	72	93	<b>.3343</b>
Revisión 2	85	67	98	<b>.3384</b>
Reseña 3	68	99	83	<b>.3384</b>
Reseña 4	88	88	74	<b>.3328</b>
Reseña 5	58	120	72	<b>.3646</b>
total	384	446	420	
PC _	.3072	.3568	.336	

Ya casi hemos terminado.  $P$  en la ecuación de Fleiss original es el promedio de los valores de  $P_i$ , por lo que lo calculamos sumando la columna  $P_i$  y dividiendo por el número de reseñas:

$$P = (.3343 + .3384 + .3384 + .3328 + .3646) / 5$$

$$= 1.7085 / 5$$

$$= \mathbf{.3417}$$

Luego calculamos  $P(e)$  sumando los cuadrados de los valores de  $P_c$ , así:

$$P(e) = .3072^2 + .3568^2 + .336^2 = .335$$

Ahora podemos finalmente conectar estos valores en la ecuación Kappa de Fleiss y calcular nuestra puntuación IAA:

$$\kappa = (.3417 - .335) / (1 - .335)$$

$$= .0067 / .665$$

$$= .010$$

### Interpretación de los coeficientes Kappa

En las dos secciones anteriores encontramos dos valores diferentes de  $\kappa$  para dos tareas de anotación diferentes, .359 y .010.

Pero, ¿cómo se interpretan realmente estas puntuaciones? En muchos casos, la interpretación del kappa depende de la complejidad y objetividad de la tarea de anotación, por lo que no existe una regla estricta que siempre se pueda usar para determinar si las puntuaciones son buenas o no. Por ejemplo, se esperaría que una tarea como el etiquetado de puntos de venta, aunque requiere mucho esfuerzo, obtenga una puntuación  $\kappa$  cercana a 1,0, debido a lo bien definidos que están los términos y las teorías subyacentes. Por otro lado, las tareas que requieren una mayor interpretación del texto, como el etiquetado de roles semánticos o la anotación temporal (donde el texto puede tener múltiples interpretaciones válidas), generalmente no tienen un estándar tan alto.

**Landis y Koch 1977** brindan estas pautas para interpretar  $\kappa$  y otras métricas de concordancia:

<b>k</b>	<b>Nivel de acuerdo</b>
----------	-------------------------

< 0	pobre
-----	-------

0,01–0,20	leve
-----------	------

0,21–0,40	justo
-----------	-------

0,41–0,60	moderado
-----------	----------

0,61–0,80	sustancial
-----------	------------

0,81–1,00	perfecto
-----------	----------



Ha habido cierto debate sobre cuán estrictamente deben interpretarse estos números, pero proporcionan un buen lugar para comenzar. Otra forma de determinar cómo deben interpretarse sus puntajes IAA es investigar un poco y averiguar qué tipo de puntajes IAA obtienen otras tareas de anotación similares. Naturalmente, siempre debe esforzarse por hacerlo bien, pero algunas tareas son simplemente más difíciles que otras.

Entonces, ¿qué podemos decir sobre las puntuaciones de la IAA a partir de los ejemplos de las secciones anteriores? De nuestro ejemplo Cohen  $\kappa$ , obtuvimos un puntaje de concordancia de .359, que de acuerdo con este gráfico es solo "regular". Y realmente, considerando las pocas categorías que se incluyeron en la tarea (cuantas más categorías incluye una tarea, más espacio hay para la ambigüedad entre las etiquetas), una puntuación de .359 es bastante mala. Mirando la tabla de anotaciones, está claro que hubo mucha confusión sobre cómo detectar una revisión neutral, por lo que definitivamente sería necesario revisar esa parte de la guía. Además, el hecho de que el Anotador B etiquetó algunas reseñas como negativas cuando el Anotador A las etiquetó como positivas, un error bastante inusual,

En cuanto a la puntuación Kappa de Fleiss... bueno, definitivamente es una que debería revisarse. Mirando el cuadro nuevamente, podemos ver que hay mucha variación en todas las columnas; de hecho, ninguna de las revisiones parece tener un sentido real de acuerdo. Hay muchos factores que pueden influir en un proyecto colaborativo, como no poder capacitar a los anotadores, no poder garantizar que los anotadores cumplan con ciertas pautas (como el idioma nativo) o simplemente una travesura en línea; véase [el Capítulo 12](#) para la discusión de algunas de las plataformas y escollos más comunes. Por supuesto, nuestro ejemplo fue inventado, pero los puntajes de acuerdo como estos definitivamente significan que necesita revisar sus pautas de anotación y probablemente también su conjunto de datos.

Sin embargo, no se desanime por los bajos puntajes de concordancia, especialmente cuando recién comienza un nuevo ciclo de MAMA. Los resultados iniciales deficientes son normales, por lo que recomendamos el uso de un pequeño conjunto de prueba para la anotación, tal como lo hará más adelante para entrenar su algoritmo ML: es la mejor manera de resolver los problemas en su tarea de anotación.

### **Cálculo de $\kappa$ en otros contextos**

Los dos cálculos de muestra que proporcionamos solo se trataban de anotadores que aplicaban un número determinado de etiquetas a un documento, lo que facilitó mucho la determinación de acuerdos porcentuales, porque el divisor para calcular porcentajes era el mismo que el número de documentos. Sin embargo, ¿qué sucede si está tratando de calcular la concordancia de las etiquetas que se aplican a las palabras/tokens en un documento, como el etiquetado de POS o el ejemplo de anotar actores,

personajes, directores, etc. en reseñas de películas que han estado discutiendo?

Para el ejemplo de etiquetado de POS, suponiendo que cada token recibe una etiqueta, entonces el conjunto de elementos que se anotan es el mismo que el número de palabras en cada documento (o en todos los documentos a la vez, si lo prefiere). Por lo tanto, su tabla de concordancia para dos anotadores se vería muy similar a la de nuestro ejemplo de reseñas de películas positivas, negativas y neutrales, solo que los números en la tabla serían mucho más grandes porque reflejarían la cantidad de palabras en el documento. o corpus, en lugar de sólo el número de documentos.

Sin embargo, ¿qué pasa con las tareas de anotación en las que *no se* anota cada token, especialmente si la anotación es bastante escasa? Digamos, por ejemplo, que le pide a dos anotadores que solo anoten los nombres y pronombres de las criaturas en "Jabberwocky" de Lewis Carroll. "Jabberwocky" tiene 216 tokens (los signos de puntuación se consideran tokens individuales), pero la mayoría de ellos no son nombres de criaturas ni pronombres. Entonces, ¿cómo se deben contar las palabras sin anotaciones? Si consideramos que "sin etiquetar" es una etiqueta, entonces podríamos terminar con una tabla de comparación que se ve así:

		B	B	B
		criatura	pronombre	sin etiquetar
A	criatura	5	0	2
A	pronombre	0	10	2
A	sin etiquetar	1	7	189

Aparentemente, nuestros anotadores imaginarios estaban un poco confundidos acerca de nuestras pautas para los pronombres y si algunas de las palabras inventadas eran criaturas u otra cosa. Independientemente, si calculamos el Kappa de Cohen para esta tabla, terminamos con un coeficiente de .691, lo cual no está nada mal considerando que la mayoría de los sustantivos del poema están compuestos de todos modos. Sin embargo, está claro que la puntuación de la IAA está muy influenciada por el hecho de que hay muy pocas etiquetas en el texto: la mayoría de los tokens no están etiquetados, por lo que el acuerdo sobre la falta de etiquetas puede estar sesgando los *resultados* .

$$Pr(a) = .944$$

$$a(criatura) = .032, b(criatura) = .028 \rightarrow criatura = .0009$$

$$a(pronombre) = .056, b(pronombre) = .079 \rightarrow pronombre = .0044$$

$$a(sin\ marcar) = .912, b(sin\ marcar) = .893 \rightarrow sin\ marcar = .814$$

$$PR(e) = .009 + .0044 + .814 = .819$$

$$k = (.944 - .819) / (1 - .819)$$

$$= .125 / .181$$

$$= .691$$

Se puede argumentar que en casos de escasa anotación, el conjunto de elementos que se examinan debe limitarse solo a los elementos que fueron etiquetados por uno o ambos anotadores. En este caso, el número de ítems que se investigan por concordancia se reduce a 27.

		B	B	B
		criatura	pronombre	sin etiquetar
A	criatura	5	0	2
A	pronombre	0	10	2
A	sin etiquetar	1	7	0

$$Pr(a) = .556$$

$$a(criatura) = .259, b(criatura) = .222 \rightarrow criatura = .057$$

$$a(pronombre) = .444, b(pronombre) = .629 \rightarrow pronombre = .279$$

$$a(sin\ marcar) = .296, b(sin\ marcar) = .148 \rightarrow sin\ marcar = .044$$

$$Pr(e) = .057 + .279 + .044 = .38$$

$$k = (.556 - .38) / (1 - .38)$$

$$= .176 / .62$$

$$= .284$$

Claramente, la puntuación de IAA cae drásticamente de un acuerdo "sustancial" en .691 a un acuerdo meramente "justo" en .284 cuando el espacio de investigación se limita solo a los tokens que fueron etiquetados por uno o más anotadores. Puede imaginar cuánto más complicada puede volverse la discusión de los puntajes de IAA cuando se evalúa una tarea que involucra anotaciones a nivel de carácter en lugar de a nivel de token, ¡donde la longitud de la palabra que se anota también puede afectar los puntajes de IAA!

Se pueden hacer argumentos convincentes para ambos lados de estas discusiones de cálculo de IAA (algunos argumentarían que, dado que las pautas de anotación también deben especificar qué *no* anotar, descartar el acuerdo en ese punto es una tontería; otros dirían que usar métricas de concordancia más rigurosas aumenta el valor de los cálculos). No estamos tomando partido sobre qué métrica es mejor, pero sea cual sea el método que utilice para estos cálculos, asegúrese de informar completamente sobre cómo calculó el acuerdo. Independientemente de cómo calcule su puntaje IAA, recuerde que indica un acuerdo relativo, y para nuestro ejemplo "Jabberwocky", ambos conjuntos de puntajes indicarían que será necesario revisar las pautas y realizar otra ronda de anotaciones, de vuelta al ciclo MAMA. !

---

#### NOTA

Si sus etiquetas tienen atributos, ¡no olvide calcular las puntuaciones IAA para ellos también! En esos casos, está bien solo comparar atributos para una etiqueta donde los anotadores están de acuerdo en que debería existir una etiqueta (ya que comparar etiquetas cuando los anotadores no están de acuerdo con la existencia de una etiqueta no le dará ninguna información útil), pero si solo compara los atributos de las etiquetas coincidentes (extensiones que son idénticas o que al menos se superponen, enlaces con los mismos anclajes, etc.), asegúrese de informar esa información con las puntuaciones.

---

## Creación del patrón oro (Adjudicación)

Una vez que haya creado un conjunto de pautas de anotación que le otorga puntajes IAA con los que está satisfecho y haya hecho que sus anotadores apliquen esas pautas a todo su corpus, es hora de adjudicar sus anotaciones y crear su conjunto de datos estándar de oro, que es lo que usará para entrenar y probar su algoritmo ML. Por lo general, es mejor tener árbitros que hayan participado en la creación de las pautas de anotación, ya que tendrán la mejor comprensión del propósito de la anotación. Contratar nuevos adjudicadores significa que tendrá el mismo problema que tuvo con los anotadores.

Dado que ya está familiarizado con la tarea de anotación, debería encontrar que el proceso de adjudicación es bastante sencillo. Necesitará software para realizar la adjudicación (vea [el Apéndice B](#) nuevamente para

obtener una lista de lo que está disponible), y después de eso, solo es cuestión de tomarse el tiempo para hacer un trabajo cuidadoso. Sin embargo, hay algunas cosas en las que pensar:

- Le tomará tanto tiempo, posiblemente más, hacer un trabajo cuidadoso de adjudicación de su corpus como le tomó a uno de sus anotadores anotarlo. No olvide asignar suficiente tiempo para hacer esto.
- Considere dividir la tarea de adjudicación en capas: primero adjudique cada etiqueta de extensión o etiqueta individualmente, luego cada etiqueta de enlace. Esto hará que sea mucho más fácil prestar atención a cómo se usa cada etiqueta y hará que las etiquetas de enlace sean mucho más precisas (porque conectarán extensiones precisas).
- ¡El hecho de que dos (o más) de sus anotadores estén de acuerdo sobre la presencia de atributos de una etiqueta en una ubicación no significa que tengan razón! Recuerde que los anotadores pueden estar de acuerdo aleatoriamente (es por eso que dedicamos tanto tiempo a calcular las puntuaciones kappa). Así que no dé por sentado que los anotadores están de acuerdo significa que tienen razón, al menos hasta que tenga una buena idea de las habilidades de sus anotadores.
- Si utiliza más de un adjudicador, considere darles a todos algunos de los mismos documentos para que también pueda calcular los puntajes de IAA para ellos; de esa manera, podrá asegurarse de que todos estén en la misma página.

Una vez que haya adjudicado todos sus archivos, estará listo para pasar a las partes de ML del ciclo MATTER.

## **Resumen**

En este capítulo, discutimos cómo aplicar su modelo y especificación a su corpus a través del proceso de anotación, y cómo crear un corpus estándar de oro que se puede usar para entrenar y probar sus algoritmos de ML. Algunos de los puntos importantes se resumen aquí:

- La "A" en el ciclo MATTER se compone de muchas partes diferentes, incluida la creación de pautas de anotación, la búsqueda de anotadores, la elección de una herramienta de anotación, la capacitación de anotadores, la verificación de puntajes IAA, la revisión de pautas y, finalmente, la adjudicación. No se desanime por la cantidad de pasos descritos en este capítulo; simplemente tómese su tiempo y tome buenas notas.
- Las pautas y las especificaciones están relacionadas, pero no son lo mismo. Las pautas determinarán cómo se aplica el modelo al texto; incluso si está utilizando el mismo modelo, si las pautas son diferentes, pueden dar como resultado un conjunto muy diferente de anotaciones.

- Crear un buen conjunto de pautas de anotación y una anotación precisa y útil no sucederá en el primer intento. Deberá revisar sus pautas y volver a capacitar a sus anotadores, probablemente más de una vez. Está bien; solo recuerde darse tiempo cuando planifique su tarea de anotación. Dentro del ciclo MATTER está el ciclo MAMA, y ninguna tarea es perfecta desde el principio.
- Una de las cosas que tendrá que considerar sobre el proceso de anotación es qué información desea presentar a sus anotadores; darles datos preprocesados podría facilitar el proceso de anotación, pero también podría sesgar a sus anotadores, así que considere qué información desea presentar a sus anotadores.
- Debido a que deberá pasar por el ciclo MAMA varias veces, es una buena idea reservar una parte de su corpus para probar sus pautas de anotación mientras resuelve los problemas. Este conjunto se puede usar en su estándar de oro más adelante, pero no debe dárselo a sus anotadores de inmediato una vez que finalicen las pautas.
- Cuando esté escribiendo sus pautas de anotación, hay algunas preguntas que necesitará responder para sus anotadores en el documento. Pero lo más importante es mantener las pautas claras y precisas, y proporcionar muchos ejemplos para que los sigan los anotadores.
- Al buscar anotadores para una tarea, debe considerar qué tipo de conocimiento necesitarán para completar su tarea de anotación con precisión (y, si es posible, rápidamente), qué idioma deben hablar de forma nativa y cuánto tiempo tiene para anotar. La última consideración puede influir en la cantidad de anotadores que necesita contratar para completar su tarea según lo programado.
- El software de anotación que le proporcione a sus anotadores para crear las anotaciones afectará la facilidad y precisión con que se crean las anotaciones, así que tenga esto en cuenta al elegir la herramienta que utilizará. El uso de más de una pieza de software para la misma tarea podría causar confusión e irregularidades en la anotación, por lo que es mejor elegir una y apegarse a ella.
- Una vez que sus anotadores hayan anotado su conjunto de textos de muestra, es hora de evaluar los puntajes de IAA. Si bien hay muchas formas diferentes de determinar el acuerdo, las dos más comunes en la lingüística computacional son Kappa de Cohen y Kappa de Fleiss. El Kappa de Cohen se usa si solo tiene dos anotadores que anotan un documento, mientras que el Kappa de Fleiss se usa para situaciones en las que se usan más de dos anotadores en un conjunto de datos.
- En función de lo buenos que sean sus puntajes de concordancia, puede decidir si su tarea está lista o no para pasar el conjunto de prueba y pasar al corpus completo. Probablemente necesitará revisar su tarea al menos una vez, así que no se desanime por los bajos puntajes de IAA.

- Interpretar las puntuaciones de IAA no es una ciencia exacta: varios factores pueden influir en si una puntuación indica que una tarea de anotación está bien definida, incluida la cantidad de etiquetas, la subjetividad de la tarea de anotación y la cantidad de anotadores.
- Además, los elementos que se anotan pueden afectar la forma en que calcula el IAA. Si bien es fácil calcular el acuerdo al aplicar una sola etiqueta a un documento completo, existe cierto debate sobre cómo se deben calcular las puntuaciones IAA al aplicar etiquetas a extensiones de texto. Independientemente del método que decida aplicar para calcular los puntajes de la IAA, realice un seguimiento de las decisiones que toma para que otras personas puedan entender cómo obtuvo sus números.
- Tener puntajes IAA altos significa que es probable que su tarea sea reproducible, lo cual es útil cuando se crea un corpus lo suficientemente grande. Sin embargo, el hecho de que una tarea sea reproducible no significa necesariamente que sea adecuada para alimentar los algoritmos de ML. Del mismo modo, el hecho de que una tarea no tenga excelentes puntajes de concordancia no significa que no sea buena para las tareas de ML. Sin embargo, será más fácil crear un corpus grande para una tarea reproducible, y cuanto más grande sea su corpus, más probabilidades tendrá de obtener buenos resultados de ML, por lo que poner un poco de esfuerzo en crear sus pautas de anotación valdrá la pena al final.
- Una vez que haya alcanzado puntajes IAA aceptables en su conjunto de pruebas de anotación, puede liberar a sus anotadores en el corpus completo. Cuando tenga el corpus completo, es hora de adjudicar las anotaciones y crear el corpus estándar de oro que usará para entrenar sus algoritmos de ML.
- La adjudicación la realizan mejor las personas que ayudaron a crear las pautas de anotación. Traer nuevas personas para realizar la adjudicación puede causar más confusión y ruido en su conjunto de datos.
- Calcular las puntuaciones de acuerdo de IAA entre los adjudicadores puede ser una buena manera de garantizar que su corpus adjudicado sea consistente. Cuanto más consistente sea su corpus, más precisos serán sus resultados de ML.