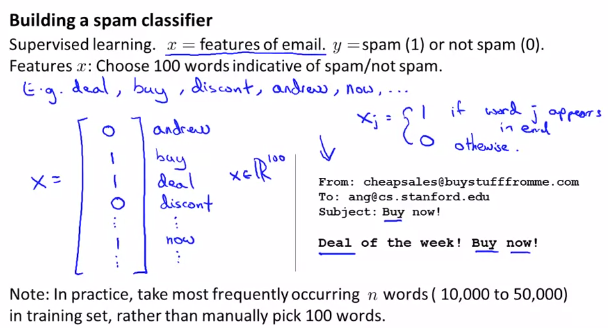
**DETECCIÓN DE CORREO SPAM**

1. Revisar la data de entregamiento y escoger las “n” palabras más frecuentes.



1. Preguntar: ¿cuál es el mejor uso de mí tiempo si quiero mejorar el uso de mi clasificador y obtener un error menor?

En el ámbito de SPAM hay proyectos llamados HONEY POT que crean direcciones falsas de correos electrónicos con el fin de lanzarlos a los correos y mejorar el entrenamiento de los algoritmos.

1. Analizar el cuerpo del texto para encontrar variables más sofisticadas.
2. Analizar los enrutamientos del correo. Desarrolle funciones sofisticadas (por ejemplo: uso de datos de encabezado de correo electrónico en correos electrónicos no deseados)

**ANALISIS DE ERRORES**

1. Si se está empezando a implementar algoritmos de machine learning se recomienda empezar con algoritmos simples para que la evidencia guíe nuestras acciones.
2. Con el algoritmo inicial trazar las curvas de aprendizaje para analizar la relación bias vs varianza y empezar a pensar en la incorporación de nuevas features. Esto evita la denominada OPTIMIZACIÓN PREMATURA DE PROGRAMACIÓN.
3. Luego de la predicción obtenida por mi algoritmo inicial, se debe tener una tarea de revisión manual de los pronósticos y de los valores reales para determinar si se pueden encontrar patrones sistemáticos para la creación de nuevas features.
4. Se debe tener una evaluación numérica del algoritmo, que a través de un número simple arroje la eficiencia el mismo.
5. Por ejemplo para problemas de detección de SPAM se puede hacer uso de lematizadores, sin embargo puede ocurrir perdida de palabras con diferente contexto (universal, universidad). En estos casos se recomienda correr 2 algoritmos, 1 lematizado y otro sin lematizar para medir la mejorar con la lematización.