El tema que ahora vamos a estudiar se llama regularización. La regularización es un procedimiento con el cual se resuelve el problema de sobreajuste (overfitting). Estudien por favor del Lecture 7.1 hasta Lecture 7.4.

También les pido instalar el paquete de Python para aprendizaje automático que se llama scikit-learn. Algunos de ustedes ya lo conocen y hasta lo usaron en los programas que yo les di de tarea hacer, lo cual fue mal porque el objetivo fue programarlo todo desde cero.

Pero ahora vamos a implementar los dos programas anteriores usando funcionalidades de scikit-learn: (1) predicción de precios de casas usando el archivo "Kc_house_data.csv" y (2) clasificación de mensajes en "spam" y "ham" usando el archivo "SMS_Spam_Corpus_big.txt".

Las funcionalidades de scikit-learn que les propongo usar son:

1. CountVectorizer (páginas 194-196 del PDF del libro "Text Analytics with Python")
2. TfidfTransformer y TfidfVectorizer (páginas 196-202 del PDF del libro "Text Analytics with Python")
3. Regresión lineal: `sklearn.linear_model.`**`LinearRegression`**, lean la documentación en https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html
4. Regresión logística: `sklearn.linear_model.`**`LogisticRegression`**, lean la documentación en https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

Envíenme por favor los dos programas (de Regresión linean y Regresión logística) usando scikit-learn el 21 de abril de 2020 (el primer martes después de las vacaciones).