


Classifying Emails for Spam Detection

• فكرة المشروع:

في المدة الأخير وتحديدًا عند ظهور الـإيميل والتعامل معه، كانت هناك ظاهرة النصب والتعدي على الخصوصية في الـإيميلات مما يجعل الوضع مزعج ولهذا تم الوصول الى فكرة فصل الـإيميلات المزجة من الهامة.

هذا المشروع يهدف إلى تطوير برنامج قادر على تصنيف رسائل البريد الإلكتروني إلى فئتين رئيسيتين: البريد العشوائي (Spam) والبريد الحقيقي (Ham). ولأن تصنيف الرسائل يدويا يصعب على الانسان ويأخذ من وقته وجهده لذلك يعتبر تطوير برنامج يصنف الـإيميلات أداة فعالة للتصدي للـإيميلات المزجة وجعل تجربة استخدام البريد الإلكتروني أكثر نظامًا.

# label	text
Labels for emails. 1 means it is spam and 0 means it is ham.	The text content of the email.
	83446 unique values
1	ounce feather bowl hummingbird opec moment alabaster valkyrie dyad bread flack desperate iambic hadr...
1	wulvob get your medircations online qnb ikud viagra escapenumber escapenumber levitra escapenumber e...
0	computer connection from cnn com wednesday escapenumber may escapenumber

Reference: <https://www.kaggle.com/datasets/purusinghvi/email-spam-classification-dataset>

• مجموعة البيانات:

سيتم استخدام مجموعة بيانات تحتوي على رسائل بريد إلكتروني

مصنفة بشكل صحيح كـ spam أو ham.

يتعين أن تشمل هذه المجموعة مجموعة متنوعة من

الرسائل لتحقيق تدريب فعال لنموذج التصنيف.

وصف للبيانات ست تتكون من عمودين:

- العمود الأول (label) ويتكون من

■ 1 تعني بريد عشوائي (Spam).

■ 0 تعني بريد غير عشوائي (Ham).

- العمود الثاني (text) يحتوي على أنواع رسائل من البريد الإلكتروني (spam/ham).

• البرمجيات التي سيتم استعمالها:

احتاج الى VSCode. ستكون بايثون لغة البرمجة الأساسية لهذا المشروع. سنستخدم مكتبات مثل scikit-Learn لمهام التعلم الآلي، سيتم إجراء استكشاف البيانات وتصورها باستخدام Pandas و matplotlib. وسيتم استعمال خوارزمية (Bayes Naive) بشكل مبدئي التي سبق دراستها في المادة.

محتوى المشروع (الخطوات التي سيتم تنفيذها):

1. تقسيم وتنظيف مجموعة البيانات.
2. انشاء (Model) لتصنيف رسائل البريد الإلكتروني.
3. حساب accuracy score وطباعة confusion matrix.
4. السماح للمستخدم بإدخال إيميل وسيقوم البرنامج بالتنبؤ بي الإيميل على أساس انه spam او لا.

• إنجاز منتصف المدة:

بحلول 25 فبراير، يتوقع أن يتم تقسيم وتنظيف مجموعة البيانات ويكون قد تم تنفيذ وتدريب النموذج بنجاح باستخدام مجموعة البيانات المختارة. سيتم تحليل النتائج التجريبية وتعديل النموذج إذا لزم الأمر لمحاولة تحسينه في التصنيف.