كلية الهندسة المعلوماتية - قسم الذكاء الصنعي عملي معالجة اللغات الطبيعية 2022-2023 الوظيفة الأولى

# التصنيف الآلي للنصوص - Text Classification

ستتعلم في هذه الوظيفة ما يلي:

- التعامل مع النصوص العربية عموماً والنصوص الطبية خصوصاً.
  - تحليل البيانات لفهم محتواها بدون قرائتها كلها.
  - تنظيف وتوحيد النص بهدف تجهيزه للقيام بعملية التصنيف.
- تصنيف النصوص باستخدام طرق مختلفة للتمثيل الشعاعي (Vectorization) وخوارزميات التصنيف الآلي (Classification).
  - دراسة تأثیر عملیات المعالجة المسبقة على دقة عملیة التصنیف.
- العمل على المكاتب الجديدة مثل pandas، nltk، sklearn وغيرها من المكاتب التي لم يسبق للطالب التعامل معها.

مرفق مع الوظيفة ملف notebook، يطلب منك أن تقوم بمئ الخلايا وفق الطلبات المدرجة أدناه. في حال طلب منك تثبيت ملاحظاتك قم بإضافة خلية نصية واكتب ملاحظاتك باللغة العربية الفصحى بطريقة سليمة. لا تنسى أن تضع شرح بسيط قبل التوابع في حال كانت معقدة وتستلزم الشرح وحافظ على نظافة الكود لتسهيل قرائته (وبالتالي تسهيل حصولك على علامة جيدة نك).

## أولاً: تحميل البيانات

تم تجميع عدد من الأسئلة والأجوبة الطبية المنمطة حسب نوع الاختصاص من موقع "الطبي", وبلغ عددها 87930 سؤال وجواب تابعة لثماني اختصاصات هي: الأورام الخبيثة والحميدة، جراحة عامة، ارتفاع ضغط الدم، جراحة العظام، أمراض الدم, أمراض الجهاز التنفسي، مرض السكري، أمراض الغدد الصماء.

قسمت الداتا إلى ثلاث مجموعات، وهي بيانات التدريب (train) و وتحقق(validation) والاختبار (test) بحيث:

- قسم التدريب train من أجل تدريب النماذج فقط.
- قسم التحقق validation ستختبر عليه نتائج النماذج التي دربتها بهدف توليف/ضبط tuning المعاملات الفائقة hyperparameter من جهة، وأيضاً للتأكد من عدم وجود مشاكل overfitting/underfitting
- قسم الاختبار test ستختبر عليه دقة النماذج النهائية بعد قيامك بضبط المعاملات، لمقارنة أداء النماذج مع بعضها وحساب قيم مقاييس الاختبار evaluation metrics.

**ملاحظة**: مجموعة البيانات التي ستستخدمها الآن هي مجموعة جزئية من مجموعة أكبر وباختصاصات أكثر ستعمل عليها في الوظائف القادمة.

يمكنك تحميل هذه البيانات من الرابط:

https://drive.google.com/drive/folders/1ETN3m1v2qhdbvOalfIA\_h-pyg48iAXy\_?usp=sharing أو استدعاء التعليمات المكتوبة في بداية الـ notebook المرفق لتحميلها لديك.

- 1. 1- قم بتحميل البيانات على data frame وطباعة أول 10 أسطر منها وعدد الداتا الكلي في كل مجموعة وحذف الأسطر المكررة أو الأسطر التي تحتوى خلايا فارغة أو تحتوى Nan.
- 1. 2-. قم بتوحيد أسماء الاختصاصات في العمود label لأن الاختصاص الواحد يمكن أن يكون مكتوب بأكثر من طريقة، مثال: جراحة-عامة, جراحة عامة، جراحة\_عامة، جراحه عامة ... و اشرح الطريقة التي اتبعتها من أجل التوحيد.

## ثانياً: فهم وتحليل النص Text Analysis

في هذه المرحلة سنقوم بمجموعة من العمليات بهدف فهم النصوص بصفوفها المختلفة لتتمكن من تنظيفها بشكل أفضل واقتراح سمات مفيدة عند التصنيف، ومن ثم تطبيق آلية التصنيف المناسبة لها. في هذه المرحلة ستعمل على قسم التدريب (train) فقط من البيانات التي لديك، وذلك لكي تتجنب تسريب البيانات (data leakage).

والطلوب:

- 2. 1- احسب عدد النصوص (سؤال/جواب) في كل اختصاص وارسم pie chart لها.
- 2. 2- قم بتقسيم النص إلى وحدات (tokenization) بطريقة مناسبة، لاحظ أن تقسيم النص على الفراغات وعلامات الترقيم يسبب مشكلة وذلك لاحتواء النص على روابط و اختصارات للألقاب ...

قم بطباعة أول 5 أسطر بعد التقسيم.

احسب عدد الكلمات الكلي (total tokens) وعدد الكلمات الفريدة الكلي (vocab) في بيانات التدريب.

2. 3- ارسم Word Cloud لكلمات مجموعة التدريب كاملةً، (انتبه أن الكلمات المكتوبة باللغة العربية تحتاج إلى إعدادات خاصة لعرضها، قم بالبحث عنها وتطبيقها في الكود الخاص بك).

اشرح في خلية نصية مختلفة استنتاجك من الرسم.

- 2. 4- قم بكتابة تابع يأخذ نصاً مقطعاً tokenized text (نص واحد وليس كل النصوص في ال data frame) ويوجد كل من:
  - أكثر 15 كلمة مكررة في النص.
  - 10 من الكلمات الأقل تكراراً في النص (hapaxes)
    - أكثر 10 ثنائيات كلمات مكررة في النص.
    - أكثر 10 ثلاثيات كلمات مكررة في النص.
  - أكثر 10 ثنائيات كلمات مهمة في النص (Bigram Collocations)
  - أكثر 10 ثلاثيات كلمات مهمة في النص (Trigram Collocations)

اشرح الطريقة التي أوجدت بها الكلمات المهمة collocations.

طبق التابع السابق على البيانات كاملة وعلى كل الاختصاص على حدى.

هل هناك كلمات أو عبارات مميزة لأي من الصفوف؟ هل ثنائيات وثلاثيات الكلمات الأكثر تكرارً هي نفسها الـ collocations؟ اكتب ملاحظاتك.

2. 5- ارسم histogram لطول النصوص بالنسبة لعدد الكلمات الفريدة في كل سؤال مدمج مع الإجابة. ثمارسمه من أجل كل اختصاص على حدى.

ماذا تستنتج؟ اكتب ملاحظتك.

في حال وجود أمر مستعصٍ عليك حاول البحث والقراءة عنه من الانترنت ولا تنتظر نزول إلهام من التأمل في سقف الغرفة الخاص بك.

## ثَالثاً: عمليات تنظيف وتوحيد النص Text Cleaning and Normalization

لا بد أنك لاحظت ;كمية المشاكل التي تحتويها البيانات وكونها خليط ما بين الفصحى والعامي، سنقوم بمجموعة من العمليات لتجهيز البيانات لعملية التصنيف، بحيث نوحد طريقة الكتابة فيها قدر الإمكان. من أجل كل من العمليات التالية قم بكتابة تابع يأخذ نصاً واحداً ويعيد النص بعد القيام بالعملية المطلوبة (تابع لكل عملية). ثم سوف تقوم باختبار التابع نفسه بطباعة خرج تطبيقه على أول 5 اسطر من مجموعة التدريب.

(قم بتحقيق هذه التوابع باستخدام التعابير النظامية فقط وليس باستبدال الحروف أو المرور عليها الواحد تلو الآخر)

- 3. 1- إزالة الروابط من النصوص، لاحظ أن بعض روابط الإجابات لا تبدأ بhttp.
  - 3. 2- حذف أسماء الأطباء مع اللقب واختصاره (د. فلان) من الإجابات.
    - 3. 3- حذف اختصاص الطبيب الموجود بعد اسمه من الإجابات.
- 3. 4- حذف المحارف المكررة من النص (مثل جدااااا ← جدا)، (لاحظ أن الكلمات العربية يمكن أن تسمح بتكرار حرفين فقط).
  - 3. 5- التعامل مع الأرقام، قم بتحقيق كل من الطرق التالية، (عند تنظيف بياناتك قد تقوم بأي من هذه العمليات، ستحدد الأفضل منها لاحقاً):
  - 3. 5- 1. توحيد رموز كتابة الأرقام، قم باستبدال الأرقام المكتوبة بالرموز الهندية ( ١ ٢ ٣ ...) بالأرقام المكتوبة بالرموز العربية (1 2 3 ...) (عادةً تدخل الباء على الأمر المراد استبداله ولكن كُتبت الصياغة بشكل خاطئ هنا لضمان عدم ضياعك عند التفكير بالمستبدل والمستبدل به 🙂 )
    - 3. 5- 2. توحيد كتابة الأرقام باستبدالها جميعها (مهما كانت رموزها) برمز ما من اختيارك.
      - 3. 5- 3. حذف جميع الأرقام (مهما كانت رموزها) من التغريدات.
        - 3. 5- 4. حذف التاريخ والوقت من الإجابات.
  - 3. 6- التعامل مع المحارف غير المرغوب بها، (عند تنظيف بياناتك قد تقوم بأي من هذه العمليات، ستحدد الأفضل منها لاحقاً):
    - 3. 6-1. حذف علامات الترقيم جميعها (عربية ولاتينية).
- 3. 6- 2. حذف جميع المحارف غير العربية (باستثناء الأرقام وعلامات الترقيم)، انتبه أنه من الأسهل لك هنا أن تفكر بما يجب أن تبقيه من النص بدلاً من التفكير بما يجب أن تحذفه، وانتبه أيضاً إلى ما تعبر عنه الكلمات غير العربية الموجودة.
  - 3. 7- حذف كلمات التوقف stop words (هل هناك كلمات توقف تظن أنه من الأفضل أن تبقيها؟). بعض الكلمات مكررة جداً وليست من كلمات التوقف مثل: إقرأ المزيد، أرجو الرد، ... الخ، قم بحذفها أيضاً.
    - 3. 8- تجذيع أو تجذير الكلمات العربية فقط.
      - 3. 9- عمليات توحيد النصوص العربية:
    - 3. 9-1. توحيد طريقة كتابة الهمزات (ء ؤ ئ).
      - 3. 9- 2. توحيد طريقة كتابة الألفات (أ آ إ)

- 3. 9- 3. حذف التطــــويل
- 3. 9- 4. حذف علامات التشكيل.
- 3. 10- احذف الفراغات المكررة، بين الكلمات واستبدلها بفراغ واحد.

#### <u>اضافی:</u>

- 3. 11- تصحيح الأخطاء الإملائية والكلمات غير الموحدة.
- بعض الكلمات وخصوصاً المعربة منها قد تُكتب كل مرة بطريقة مختلفة، مثل:أسماء الأدوية وغيرها وقد يكون التجذيع والتجذير غير مفيد هنا في توحيدها، اكتب تابع لتوحيد الكلمات التي يمكن كتابتها بطرق مختلفة.
- لاحظ أن التغريدات مكتوبة بمزيج من اللهجات والفصحى، ولاحظ أيضاً مقدار الأخطاء الإملائية الموجودة فيها، مثل: (قسطرة وقثطرة وقصطرة، لكن ولاكن، ظغط وضغط، ... الخ) اكتب تابع لتصحيح هذه الأخطاء أو توحيدها.
- بعض الكلمات تكتب سهواً بدون فراغ بينها، مثل: صياممتقطع، ... أوجد طريقة لفصلها. اشرح الخوارزمية التي نفذتها لحل الطلب في خلية منفصلة.

**ملاحظة هامة**: المعلومات التي يمكن أن تستخدمها لحل هذا الطلب لو وجدت (مثل تكرار الكلمات) لا تحصل عليها إلا من مجموعة التدريب.

أى عملية قد تجدها مفيدة في هذه المرحلة ولم تذكر بإمكانك تطبيقها مع شرح رؤيتك لفائدتها ...

## رابعاً: تجهيز بيانات التدريب Prepare Training Data

الآن سنطبق العمليات السابقة على البيانات لتصبح جاهزة للمرحلة التالية من الوظيفة. تذكر أن هذه العمليات ليست بالضرورة مفيدة لعملية التصنيف، ولا يمكن معرفة ذلك إلا بالتجريب.

- 4. 1- قم بكتابة تابع عام يأخذ قائمة من التوابع المرتبة حسب ترتيب تنفيذها بحيث لا تعدل التوابع السابقة أو هذا التابع عند كل تطبيق لضمان كتابة كود جيد ونظيف، واختبر التابع على الأسطر الخمسة الأولى في بيانات التدريب.
- 4. 2- طبق التابع السابق بوجود العمليات من المرحلة السابقة على مجموعة التدريب كاملةً، واحسب عدد الكلمات الكلية والفريدة قبل التنظيف. ثبت ملاحظتك.
- 4. 3- هل يوجد خلايا أصبحت فارغة أو مكونة من عدد محارف قليل (أقل من 5) بعد التنظيف؟ قم بحذف هذه الأسطر إن وجدت، ما عددها؟

ستعمل عند التصنيف على السؤال والجواب معاً أي يجب أن تقوم بدمجهما concatination في نص واحد لكل سطر، وكذلك على تصنيف السؤال لوحده (الحالة الحقيقية في التطبيق الطبي حيث يصنف السؤال قبل الإجابة عنه لتوجيهه إلى القسم أو الطبيب المختص).

## خامساً: النموذج الحدودي (Baseline)

من أجل كل عملية تجريب ستقوم بها قم بطباعة الدقة، ووضع رقم الطلب واسم النموذج ومعاملاته وعمليات المعالجة مع الدقة في ال dictionary الذي ستستخدمه في نهاية كل طلب لطبع جداول المقارنة بين كل النماذج حسب المثال الموجود في الـ notebook المرفق وحسب الإرشادات في آخر الوظيفة. بالإضافة إلى ذلك سيكون هناك جدول نهائي ستثبت فيه نتائج أفضل عملية تجريب قمت بها بكل طلب.

- 5. 1- من أجل كل مجموعة من مجموعات البيانات الثلاث استخرج مصفوفة الدخل (X) شعاع الهدف (y)الممثل بالاختصاصات (multi-class classification).
- 5. 2- طبق خوارزمية التصنيف multinomial Logistic Regression على مجموعة التدريب (خوارزمية logistic بتابع sigmoid بتابع softmax )، وذلك بعد تمثيل regression ولكن تستطيع تصنيف لأكثر من صنفين باستبدال softmax بتابع bag of Words )، وذلك بعد تمثيل البيانات (السؤال والجواب) بطريقة Bag of Words، لا تقم بأي عملية تنظيف على البيانات بل استخدمها بشكلها الخام.

اختبر النموذج على مجموعة الاختبار، قم بتثبيت النتيجة في جدول النتائج النهائية كما هو موضح في المثال في المثال في المثال في notebook المرفق. لاحظ أن البيانات غير متوازنة imbalance data في الصفوف وهذا قد ينتج دقة غير صحيحة تماماً، استخدم المقياس balance accuracy من مكتبة sklearn واضفه إلى الجدول. ثبت ملاحظتك عن الفرق بين المقياسين.

(يفضل أن تكتب الكود السابق في تابع لأنك ستقوم باستدعائه أكثر من مرة من أجل الطلب اللاحق.)

- 5. 3- قم بتدريب النموذج السابق باستخدام عملية تنظيف أو توحيد واحدة كل مرة واحسب الدقات و اطبعها ثم ادرجها في جدول النتائج لهذا الطلب، وفي حال كانت العملية لم تحسن النتيجة لن تقوم باستخدامها في الطلب التالى.
- 5. 4- درب النموذج على مجموعة العمليات التي حسنت نتيجة الاختبار بالمقارنة مع النموذج المدرب في الطلب(5. 1) واختباره على مجموعة الاختبار ثم اطبع الدقات وثبتها في الجدول.
- 5. 5- باستخدام أفضل عمليات التنظيف والتوحيد الناتجة عن الخطوة السابقة. قم بتمثيل البيانات باستخدام نموذج TF-IDF مع logistic regression، بمعاملاته الافتراضية، اختبر نموذجك وثبت النتائج في جدول نتائج الطلب.
  - 5. 6- قم بضبط معاملات TF-IDF على مجموعة التحقق validation set للحصول على أفضل نتيجة، وثبت نتيجة الاختبارات في جدول نتائج الطلب.

بعد الانتهاء قم باختبار النموذج الأفضل على بيانات الاختبار، ثبت النتائج في جدول الاختبار النهائي.

مجموعة العمليات التي أعطتك أفضل نتائج عند تطبيق عملية تمثيل البيانات باستخدام نموذج Bag of words لن تعطي بالضرورة أفضل نتائج عند تطبيق نموذج TF-IDF، وأيضاً تختلف باختلاف خوارزمية التدريب وحتى معاملات كل نموذج ولكن للتسهيل لن نقوم باختبار ذلك هنا.

### سادساً: تمثيل الدلالة Semantic Representation

ستعمل في هذا القسم على إيجاد تمثيل للكلمات وإيجاد العلاقات بينها بناءً على هذا التمثيل.

6. 1. اكتب تابع يستخدم مصفوفة bag of words السابقة للحصول على شعاع التكرارات الخام لكلمة معينة من أجل كل نصوص مجموعة التدريب، واطبع كلمة معينة مع شعاعها باستخدام هذا التابع. 6. 2. اكتب تابع يأخذ كلمة ما ويعيد مجموعة من الكلمات التي يتكرر ظهورها في سياق هذه الكلمة، لاحظ أننا
 لا نتحدث عن التشابه الدلالي بين الكلمات وليس بالضرورة أن ترد كلمات الخرج مباشرة قبل أو بعد كلمة
 الدخل وإنما في نفس النصوص.

قد تحتاج إلى القليل من التفكير والبحث لحل هذا الطلب.

ماذا تستنتج؟ ثبت ملاحظاتك.

- 6. قم بتصميم شبكة word2vec من نوع skipgram باستخدام مكتبة keras ودربها على نصوص مجموعة التدريب من أجل الحصول بعد انتهاء التدريب على مصفوفة أشعة الكلمات المضمنة embedding vectors.
  في هذا الطلب عليك أن تفكر بما يلي:
  - هل تحتاج إلى التجذيع أو التجذير من أجل التنبؤ الكلمات الواقعة في سياق كلمة الدخل؟ ولماذا؟
    - هل تحتاج إلى عملية تقطيع مختلفة عما نفذته في الطلبات السابقة ولماذا؟
    - هل جميع الكلمات يجب أن نوجد أشعة دلالية لها؟ ما فائدة وسلبيات هذه الفكرة؟ ثبت ملاحظاتك.
- 6. 4- قم بعملية إظهار (Visualization) لمصفوفة الأشعة ( الـ embedding) الناتجة عن عملية التدريب، ثبت ملاحظاتك.
  - 6. 5- اكتب تابع يطبق عمليات anology على أشعة كلمات دخل التابع، وقم بطباعة خرجه من أجل 5
    استدعاءات له لمجموعة كلمات من اختيارك. هل ترى أن النموذج جيد؟ اكتب ملاحظتك.

# سابعاً: التدريب باستخدام شبكة عصبونية عميقة Deep Neural Network

7. 1- قم بتدريب شبكة عصبونية عميقة مؤلفة من عدة طبقات متصلة بشكل كامل (fully connected) على النصوص النظيفة بأفضل مجموعة من العمليات، وثبت النتائج في جدول النتائج. السمات التي ستسخدمها:

7. 1-1. Bag of words

7. 1- 2.TF-IDF

- 7. 2- قم بتدریب شبکة عصبونیة عمیقة من النوع (Convolutional Neural Networks (CNN) علی النصوص النظیفة، بحیث تستخدم embedding layer بأربع طرق:
- 7. 2- 1. تستخدم مصفوفة أشعة الكلمات الناتجة عن تدريب نموذج word2vec في الطلب السابق وبدون تعديلها (جمد طبقة ال embedding عند تدريب الشبكة).
- 7. 2- 2. تستخدم مصفوفة أشعة الكلمات الناتجة عن تدريب نموذج word2vec في الطلب السابق مع السماح بتعديلها أثناء التدريب.
- 7. 2- 3. لا تستخدم أشعة مسبقة لتهيئة الطبقة بل دعها تتعدل أثناء التدريب مثلها مثل أية طبقة أخرى.
- 7. 2- 4. حمل مصفوفة embedding ناتجة عن تدريب مسبق من الانترنت ومهما كانت طريقة ال embedding بدون السماح embedding بدون السماح بتعديلها أثناء التدريب.

وثبت نتائج كل طريقة في جدول نتائج الطلب بعد اختبار النموذج الطلب. ماذا تلاحظ؟ ثبت ملاحظاتك.

7. 3- لا تنسى ضبط معاملات النموذج الفائقة على مجموعة التحقق validation set بعد اختيار أفضل مصفوفة أشعة دلالية للكلمات من تجريب الطرق الأربع في الطلب السابق، واطبع مقدار مقياس التقييم. لا تنسى طباعة منحنيات التعلم لاكتشاف جودة التدريب لكل شبكة (validation) (validation)

7. 4- (**إضافى**) قم بطباعة مصفوفة التعارضات Confusion Matrix ثبت ملاحظاتك.

(**إضافي**) استخدم عند تدريب الشبكات السابقة معيار تقييم metric خاص custom function بدلاً من الدقة accuracy لأخذ عدم توازن البيانات بعين الاعتبار.

ثبت النتائج في جدول الاختبار النهائي.

ملاحظة: المعطيات ليست متوازنة imbalanced بمعنى أن عدد النصوص في كل صنف ليس متساوي، هذا قد يشكل مشكلة -ولكن ليس بالضرورة-، في حال أردت موازنتها بإمكانك ذلك، ابحث عن هذا الموضوع (resampling) أو قم بموازنتها يدوياً عن طريق أخذ عدد متساوي من النصوص بشكل عشوائي لكل صنف، سنتكلم عن هذا الموضوع بشكل مستفيض في مادة التعلم الآلي.

### الإرشادات

- تسلم الوظيفة قبل يوم الثلاثاء 5-12-2013 الساعة 59:11 مساًء. يمكنك تسليم الوظيفة متأخرً ولكن سوف يترتب على ذلك حذف جزء من العلامة (10% من العلامة في حال سلمت خلال الساعات الـ 12 الأولى من انتهاء الموعد، ومن ثم 5% عن كل يوم تأخير).
- الوظيفة تتطلب مجموعة مكونة من 3 طلاب لا أقل ولا أكثر ولا مجال للجدال في العدد المسموح، وغير مسموح بتغيير أفراد المجموعة في الوظائف القادمة، تحتاج الوظيفة لتنتهي في الوقت المحدد تعاون الأفراد الثلاثة (الوقت محسوب بدقة)، كما أن علامة الطلاب في المجموعة الواحدة ليست موحدة، يحصل الطلاب على جزء من العلامة على حل الوظيفة والجزء الآخر على مقدار عمل الطالب بها، معرفة عمل كل طالب ومقداره ستتم بالتأكيد بطريقة غير معلنة الآن والاختيار لك:

#### "live or die, make your choice. The game has just begun"

قم بتغيير اسم الملف بكتابة اسمك باللغة العربية مكان [your\_name]، قم برفع نسختين من الـ notebook الأول بصيغة ipynb والثاني بصيغة html وملف جدول النتائج النهائية على الرابط الآتي يدون ضغطه:

#### https://forms.gle/VMdbgHuZAviTY2oj8

- عندما تقوم باختبار الكود الخاص بك تأكد من أنك لم تقم بطباعة كل البيانات ضمن النوتبوك يكفي أن تطبع حالات الاختبار المطلوبة منك.
- تأكد قبل تسليمك للملف أن جميع الخلايا منفذة بشكل كامل والنتائج معروضة فيه. وتأكد أن الملف يعمل وأنه قابل للقراءة بوضوح، قم بفتح ملف الـ html وتأكد أنه صحيح وأن حجم الملف المسلم

- صغير، في حال كان كبيراً تأكد من أنك لم تقم بطباعة كل البيانات بشكل غير مقصود فيه. في حال تجاوز حجمه الحد المسموح (10M) لن يتم رفعه.
  - لن تقبل الوظيفة بدون طباعة جدول مقارنة النماذج.
- نتائج كل طلب والنتائج النهائية تخزن في dictionary يتألف من 7 مفاتيح، الأول باسم model\_name ويعبر عن question\_step\_number ويعبر عن رقم السؤال ورقم الطلب، والثاني باسم question\_step\_number ويعبر عن شكل الدخل أو سمات النموذج، والرابع باسم features ويعبر عن شكل الدخل أو سمات النموذج، والرابع باسم model\_parameters وقيمها الجديدة وفي حال عدم تعديل بارامترات النموذج تضاف كلمة default، والخامس preprocessing\_methods وفي حال عدم تعديل بارامترات النموذج تضاف كلمة التدريب، والسادس باسم accuracy ويعبر عن مقدار دقة عدم التوازن على بيانات دقة الاختبار، وأخيراً السابع باسم balance\_accuracy ويعبر عن مقدار دقة عدم التوازن على بيانات الاختبار. كل مفتاح منها قيمته الانتراكة الدخلها بتعليمة append
  - اختر لأسماء النماذج وعمليات المعالجة أسماءً واضحة ومعبرة، في حال عدم وجود أي إضافة قم ىكتابة none.
- انتبه أنه هناك dictionary يضاف إليه نتائج طلبات التدريب، وهناك جدول نهائي للنماذج الأفضل من كل طلب منها.

#### وتذكر أن:

تنفذ الإرشادات وتعليمات التسليم حرفياً دون إبداعات إضافية، لا زيادة ولا نقصان. عند وجود أي تشابه بين وظيفتي مجموعتين ستخسر المجموعتان العلامة معاً دون مراجعتهما (هذا خبر وليس تهديد عزيزي الطالب).

مدرسو المادة: زينة الدلال، عُلا طبال، ايليسار بري