Projects Guidelines

م. زينة الدلال م. علا طبال م. ايليسار بري

أين يمكن أن أجد أوراق بحثية؟

الأفضل دائماً أن تبدأ بمراجعة ودراسة الأعمال السابقة المشابهة لعملك من حيث الهدف أو التقنية المستخدمة أو حتى مجموعة البيانات، إذ يجنبك ذلك الوقوع في الكثير من الأخطاء والتجارب الفاشلة أثناء العمل كما أنه قد يوجهك نحو نهج ما كانت لتخطر على بالك بدون هذه الدراسة، بالإضافة إلى أنها تسمح لك تقييم عملك ومقارنته مع أعمال الدراسة، وبهذا تكون قد حددت أين نحن وما الذي يمكن أن نفعله.
إذاً الدراسة المرجعية توفير للوقت والجهد فيما بعد.
يمكنك أن تجد أفضل الأعمال في مجال الذكاء الصنعي بشكل عام ومعالجة اللغات الطبيعية بشكل خاص المنشورة في أو من قبل الدوريات journals والمؤتمرات conferences التي تنشر وتقدم هذه الأوراق البحثية بعد مراجعتها والتأكد من صلاحيتها العلمية إن صح التعبير، وأهمها:

- Association for Computational Linguistics - ACL
- Empirical Methods in Natural Language Processing - EMNLP
- Transactions of the Association for Computational Linguistics - TACL
- The North American Chapter of the Association for Computational Linguistics - NAACL
- the European Chapter of the Association for Computational Linguistics - EACL
- Neural Information Processing Systems - NIPS
- The International Conference on Learning Representations - ICLR
- The International Conference on Machine Learning - ICML
- Arabic Natural Language Processing Workshop - WANLP
- International Conference on Natural Language Computing and Semantic Processing (ICNLCSP)
- International Conference on Natural Language Processing and Transition Networks (ICNLPTN)
- Journal of King Saud University - Computer and Information Sciences.
- The Association For The Advancement Of Artificial Intelligence - AAAI

يمكنك إيجاد هذه الأوراق أو الأعمال السابقة في:

- http://arxiv.org/ or http://www.arxiv-sanity.com
- https://papers.nips.cc
- http://scholar.google.com
- http://dl.acm.org/
- https://nlp.stanford.edu/pubs/
- http://aclasb.dfki.de/

- https://paperswithcode.com/sota
- https://nlpprogress.com/

يمكنك أيضاً استخدام الأداة https://huyenchip.com/2018/10/04/sotawhat.html التي تتيح لك كل الأوراق التي ذُكرت فيها كلمة معينة، من الأدوات المفيدة أيضاً:

https://www.researchrabbit.ai/

كيف ابحث؟

عملية البحث عن أعمال سابقة عملية تكرارية، فقد لا تمتلك في البداية الكلمات المفتاحية keywords المناسبة للبحث عن الموضوع فتقوم باستخدام ما يخطر ببالك من كلمات لتحصل على أوراق وأعمال قد لا تكون مناسبة أو ذات صلة ولكن من خلال الاطلاع عليها قد تجمع المزيد من الكلمات المفتاحية وتكرر البحث حتى تحصل على الكلمات المناسبة وبالتالي الأوراق والمقالات المناسبة.

أي معلومة تحتاجها موجودة حتماً على الإنترنت، لكن الوقت اللازم للوصول إليها هو ما يميز الخبراء عن المبتدئين، حتى تصبح خبيراً في إيجاد أي معلومة لابد أن تتعلم أساسيات البحث وأولها إيجاد الكلمات المفتاحية كما ذكرنا سابقاً، وتتعلم ما يسمى Search operators، وأهمها:

- OR
- Site : توجيه محرك البحث للبحث ضمن موقع أو domain معين، على سبيل المثال: sy. توجيه المحرك للبحث ضمن المواقع السورية فقط.
- Filetype : توجيه المحرك للبحث عن نوع معين الملفات
- إشارة - : لحذف النتائج التي تحتوي الكلمات بعد هذه الإشارة
- إشارات التنصيص "" : لتوجيه المحرك للبحث عن الحملة الموجودة ضمن إشارات التنصيص بشكل حرفي وبالتالي حساسة للأخطاء الإملائية.

يمكن استخدام المزيد من هذه العمليات في واجهة البحث المتقدم، وللمزيد من الأمثلة والشروحات عن هذا الموضوع يمكنك مشاهدة هذه الفيديوهات الممتعة:

https://www.youtube.com/watch?v=erZ3IyBCXdY

https://www.youtube.com/watch?v=hmRakMppVtw

"*Stop searching like my grandma*"

← ماذا تفعل في حال كانت الأبحاث غير متاحة للتحميل أو وجدت بعض الكتب المفيدة لكنها مدفوعة؟

قم بالاستعانة بهذا الموقع حيث يكفي أن تضع فيه رابط الكتاب/ الورقة البحثية أو المعرف الخاص بها doi لتحصل عليها مجاناً:

https://sci-hub.st/

كما يمكنك إيجاد الكثير من الكتب المدفوعة مجاناً وحتى تلك الموجودة على أمازون في هذا الموقع (من النادر أن تجد فيه كتب عربية مع الأسف):

http://libgen.rs/

وكذلك موقع https://z-lib.org/ إلا أنه محدود بعدد التحميلات المتاحة للحساب كل يوم.
مشكلة هذه المواقع هو تغير الـ domain المستمر، في حال قمت بفتحها ولم تعمل قم بالبحث عن الدومين الجديد لها.

كيف أقرأ ورقة بحثية؟

بالطبع ليست كل ورقة بحثية تجدها قد تكون مفيدة لك، أو حتى مفيدة بشكل عام والسبب في ذلك هو فقاعات ما يدعى hype، الناتجة عن انبهار عامة الناس في مجال الذكاء الصنعي وبالتالي استقطاب المزيد من الأشخاص إليه وبالتالي المزيد من الأبحاث، وهو أمر جيد ولكن من آثاره السلبية أن يجعل الناس تستخدمه فقط من أجل الجذب، كاستخدامه في كتابة مقالات بعناوين لافتة من قبل الصحفيين لجذب القراء ومن قبل الشركات لجذب المستثمرين وحتى من قبل الباحثين لجذب المزيد من الاستشهادات citations بأوراقهم البحثية حتى وإن كان جوهر الفكرة لا يستحق، لذلك من المهم أن تظل على دراية بالدوافع المختلفة وراء ما يُكتب.
أحد أكثر مظاهر الhype شيوعاً هو إعادة تسمية الأفكار القديمة بتسميات جديدة براقة وطنانة buzz words (مثل عبارة data science)، والمفترض الحكم على العمل من خلال التجارب التي أجريت فيه والنتائج التي خلص إليها لا العناوين.
وتمييز المفيد عن غير المفيد أمر يحتاج إلى خبرة وكذلك القدرة على التراجع عندما تصل لنهايات محسومة بدلاً من إضاعة المزيد من الوقت في أمر غير مجدي (راجع مغالطة التكلفة الغارقة sunk cost fallacy)، يقول العالم ريتشارد فاينمان:

"We are trying to prove ourselves wrong as quickly as possible, because only in that way can we find progress."

الآن حتى تبدأ بمراكمة هذه الخبرة اتبع الخطوات الآتية:

1. ضع كل ما وجدته من أبحاث وأوراق (بدون قراءتها) في مجلد خاص وبهذا تكون قمت بإعداد قائمة الأبحاث.
2. من أجل كل ورقة قم باتباع طريقة multiple passes في قراءتها، إذ أن قراءة الورقة مباشرة من البداية إلى النهاية هي أسوأ طريقة للقراءة، والمراحل المختلفة في طريقة multiple passes هي:

- تقرأ كل من العنوان والملخص abstract والمخططات والجداول figures لأنها تلخيص لكل المقالة، وفي هذه المرحلة يمكنك تحديد هل هذه الورقة مفيدة لك أم لا، في حال كانت غير مفيدة قم بشطبها من القائمة.
- تقرأ العنوان والملخص والمقدمة introduction والخلاصة conclusion بالإضافة إلى المخططات والأشكال.
- تقرأها كلها من البداية للنهاية بدون المعادلات الرياضية الموجودة فيها (skip the math).
- اقرأها كلها من جديد مع المعادلات ولكن في هذه المرحلة وبعد قراءاتك المتعددة تكون قد كونت فكرة واضحة عما هو مفيد حقاً وعما لا يضيف للورقة شيء وبالتالي هنا عند قراءتها كلها يمكنك تجنب الكلام الفارغ فيها.
- من أجل تطوير مهاراتك أكثر (إن أردت) وبعد قراءتها بشكل جيد وفهمها، ضع النتائج والملاحظات أمامك وحاول إعادة بناء المعادلات من الصفر from scratch لوحدك كما يحاول

طلاب الفنون تقليد لوحات كبار الرسامين عند التعلم، وكذلك على مستوى الكود، حاول تنفيذ الورقة بشكل مصغر لوحدك.

"*What I cannot create, I do not understand.*"

3. بعد قراءة الورقة يمكن أن تجد في أعمالها السابقة related works أبحاث مفيدة أيضاً وغير موجودة في قائمتك، قم بإضافتها إلى القائمة.
4. اطرح على نفسك الأسئلة الآتية للتأكد من فهمك:
   - ما الذي حاول الكاتب حله؟
   - ما هي العناصر الرئيسة للعمل key elements؟
   - ما المفيد من الورقة حقاً؟

تستغرق هذه العملية من أجل كل ورقة تقريباً من 5 إلى 6 ساعات بالنسبة لشخص مبتدئ، ومن ساعة إلى ساعتين للخبير، ويعتمد الوقت أيضاً على صعوبة الموضوع أو الورقة، حيث أن أخذ الوقت الكافي لهضم محتوى الورقة لا يقل أهمية ابداً عن قراءة أكبر عدد ممكن لك من الأوراق، قم بتدوين الملاحظات وتحديد الجمل الهامة بعد القراءة مباشرة حتى لا تنسى ما قراءته وتضطر فيما بعد إلى إعادة القراءة من جديد.
يحتاج تعلم أساسيات موضوع ما قراءة من 5 إلى 20 ورقة بحثية وعمل، وهذا ليس بالضرورة أن يؤدي إلى إمكانية تنفيذ implementation جيد عند بدء العمل، بينما يحتاج الأمر إلى قراءة مئات الأبحاث بالنسبة للباحثين والأشخاص الذين يريدون التخصص في مجال معين.

https://web.stanford.edu/class/ee384m/Handouts/HowtoReadPaper.pdf

هذا لا يعني أن تقوم بقراءة 20 ورقة في يومين متتالين من أجل أن تكتسب معرفة وخبرة أكبر، على العكس، قيامك بالتكثيف والانقطاع لن يكسبك شيء لأن عملية اكتساب الخبرة وتطوير المهارات يحتاج إلى التدريب والاستمرارية، الأفضل قراءة ورقتين بحثيتين كل أسبوع وبشكل مستمر، ستكتشف بعد سنة من المداومة على القراءة بهذه الطريقة مدى تطور مهاراتك في القراءة والفهم ومقدار الأشياء التي تعلمتها.
لا تتوقع أن تفهم كل ما ورد في الورقة وخصوصاً إذا كان الموضوع جديد بالنسبة لك، من أجل ذلك استعن بالفيديوهات التي تشرح المصطلح أو الفكرة وكذلك المقالات blogs الموجودة على مواقع مشابهة ل medium، وحتى يمكنك القيام بهذه الخطوة قبل البدء بأي شيء لتكوّن صورة عن الموضوع وعن بعض المصطلحات التي قد ترد في الأوراق البحثية.
في حال وجدت خبير في هذا الموضوع يستطيع إجابتك على أسئلتك مهما بلغت سذاجتها ودون أن يحكم عليك أو ينظر إليك بنظرة مختلفة فإنك قد وجدت كنزاً! لا تفرط بهذا الشخص واستمع إلى نصائحه، فحوار واحد مع شخص خبير من هذا النوع يغنيك عن الكثير من البحث غير المجدي.

كيف اكتب الدراسة المرجعية؟

عند كتابة الدراسة يجب مُراعاة عنصر الاختصار والإيجاز قدر الإمكان، ولكن يجب أن يكون ذلك بشكل غير مُخلٍّ، إذ يجب أن توضح كل النقاط الهامة في العمل كآلية الحل والنتائج التي جرى الوصول إليها بشكل مكثف.
يجب توثيق كل الأعمال المكتوبة وكل ما جرى الاستشهاد به في الدراسة، والتوثيق يكون في: متن الدراسة المرجعية وفي صفحة المراجع، ويجب أن تتطابق المراجع الموثقة في المتن مع المراجع الموثقة في صفحة المراجع.

← كيف أقوم بالتوثيق؟

التوثيق في المتن:

- لتوثيق مرجع لأول مرة ولكاتب واحد: يجب كتابة الاسم الأخير للباحث مع تاريخ النشر بين قوسين، مثال: قام الخليلي (1987) بدراسة حدد فيها ...
- لتوثيق مرجع لمؤلفين اثنين: يجب ذكر الأسماء الأخيرة للمؤلفين في كل مرة يذكر فيها المرجع مع تاريخ النشر بين قوسين، مثال: قام الخليلي وملكاوي (1985) بدراسة حول ...
  مثال باللغة الإنكليزية:

  Hall and McCurdy (1990) ...
- لتوثيق مرجع لثلاثة مؤلفين أو أكثر: يجب ذكر الأسماء الأخيرة لجميع المؤلفين عند ذكر المرجع لأول مرة ثم تكتفي بذكر الاسم الأخير للمؤلف الأول تليه كلمة (وآخرون) طبعاً مع تاريخ النشر بين قوسين كل مرة، مثال: قام الخليلي وآخرون (1985) ...
  مثال باللغة الإنكليزية:

  Butzo et al. (1997) ...
- في حال جرى الاقتباس من عمل حرفياً يوضع الاقتباس ضمن إشارتي التنصيص "" ويذكر رقم الصفحة التي تحتوي الجملة المقتبسة، مثال:
  أشار عوده وملكاوي (1987، ص 75) أن "البحث التطبيقي يشترك مع البحث ألساسي في تطبيق المنهج العلمي في البحث، إال أن هدفه ألساسي هو تحسين الواقع العلمي من خالل اختبار النظريات في مواقف حقيقية"
- لتوثيق عمل مترجم تتبع نفس الإرشادات السابقة لكن مع ذكر تاريخ الترجمة بعد تاريخ نشر العمل الأصلي في القوسين، مثال: ذكر كرونشكانك (1971 / 1993).
- لتوثيق أكثر من عمل معاً توضع المراجع بين قوسين وترتب ترتيباً أبجدياً على أسماء المؤلفين، مثال:
  (Balda, 1980; Kamil, 1988; Pepperberg & Funk, 1990)

عند كتابة المراجع في قائمة المراجع لا تكتفي بوضع رابط فقط، وتذكر أن المقالات blogs والفيديوهات ليست بمراجع، بشكل عام توثق الأعمال في قائمة المراجع كما يلي:

اسم المؤلف - اسم المرجع - مكان النشر - دار النشر - سنة النشر - الجزء - الصفحة.

- في حالة إذا كان الكتاب مترجمًا تتم كتابة المرجع على هذا النسق:

اسم المؤلف - اسم المرجع - اسم المترجم - مكان النشر - دار النشر - سنة النشر - الجزء - الصفحة.

- إذا كان المرجع دوريةً journal فتتم كتابتها على هذا النسق:

اسم المؤلف - عنوان المقالة - عنوان الدورية - رقم العدد الخاص بالمجلد - تاريخ الصدور - الصفحة.

- إذا كان المرجع عبارة عن بحث مقدم لمؤتمرات علمية:

  اسم المؤلف - عنوان البحث - موضوع المؤتمر - مكان انعقاد المؤتمر - تاريخ انعقاده.

- إذا كان المرجع عبارة عن موقع إلكتروني:

  اسم الموقع - اليوم - الشهر - السنة.

وترتب المراجع في قائمة المراجع كما يلي:

- ترتب المراجع ترتيباً أبجدياً.
- ترتب المراجع لمؤلف واحد وفقاً لسنة النشر، فالمرجع الأقدم يوضع قبل المرجع الأحدث.

- المرجع الفردي للمؤلف يوضع قبل المرجع المشترك.
- ترتب مراجع نفس المؤلف بعمل فردي صدرت في نفس العام حسب ترتيب الحرف الأول لعنوان المقالة. وإذا وجد تشابه بين الحروف الأولى في العناوين ينظر إلى الحروف التي تليها وهكذا.
- ترتب المراجع التي تتشابه في اسم الباحث الأول وتختلف في اسم الباحث الثاني وفقا للترتيب الأبجدي للحرف الأول من الاسم الأخير للباحث الثاني وهكذا.
- ترتيب المراجع لنفس المؤلفين حسب تاريخ المرجع.

لا تنس أن المقالات، الروابط، الفيديوهات وكل ما شابهها ليست مراجع (المرجع يختلف عن المصدر)، حتى أن الأوراق البحثية الموجودة ليست جميعها مؤهلة لتكون مرجع (الأمر يتعلق بتصنيف الباحث والمجلة أو المؤتمر الذي نشرها والكثير من الأمور لا تهم طلاب تحت التخرج)
للمزيد من الأمثلة عن كيفية التوثيق راجع مصدر الدكتورة ندى غنيم الموجودة في قائمة المصادر في الأسفل.

أين يمكن أن أجد فكرة؟

لتجد فكرة تعمل عليها البعض ينصح بالاطلاع على ما يجري في المجتمع العلمي والتوجهات الحالية والمشاكل أو المواضيع المثيرة hot topics حالياً، كما يقول لاعب ومدرب رياضة الهوكي الثلجية:

"*I skate to where the puck is going, not where it has been*".

وعلى الخلاف ينصح البعض الآخر بعدم الانجراف وراء الترينذ العلمي الحالي، يقول العالم Herb Simon الحائز على جائزة نوبل وجائزة تورينغ turing لطلابه:

"*If you see a research area where many people are working, go somewhere else.*"

وتذكر دائماً أن أفضل مكان لتجد فكرة رائعة هو العالم حولك حيث توجد المشاكل الحقيقية.

➔ **كيف تجد المواضيع المثيرة الحالية في مجال الذكاء الصنعي؟**

بالطبع قراءة الأبحاث الدائم والاطلاع على المؤتمرات العلمية في هذا المجال مفيد، لكن من الأمور الأبسط والمفيدة أيضاً هو متابعة الخبراء والعاملين في هذا المجال على مجموعات ووسائل التواصل الاجتماعي وبالتالي استخدامها لأمور مفيدة (على عكس ما يشاع بين الكثير من الناس عن الترويج لإغلاق حساباتك على هذه المواقع كونها مضيعة للوقت)، بالإضافة إلى النقاش مع الزملاء folks في هذا المجال (وهذا ما ينصح به andrew ng طلابه)، إذاً إلى جانب قراءة الأبحاث لمعرفة التوجهات الحالية في الذكاء الصنعي يمكنك:

- متابعة الخبراء في هذا المجال على الفيسبوك والتويتر، مثل: yann lecun, andrew ng, yoshua bengio, ...
- الانضمام إلى المجموعات التي تناقش مواضيع الذكاء الصنعي على reddit وفيسبوك، مثل:

https://www.reddit.com/r/MachineLearning/
https://www.facebook.com/groups/Arabic.A.I.Paper/
https://www.facebook.com/groups/hasoubawayh/
https://www.facebook.com/groups/555836078661044/
https://www.facebook.com/groups/nlpcrackers/
https://www.facebook.com/ai.bilarabic
https://www.facebook.com/groups/DeepLearnng/
https://www.facebook.com/groups/1395969327144025/

https://www.facebook.com/groups/2178368982486875/

https://www.facebook.com/groups/2059467967664033/

https://www.facebook.com/groups/MontrealAI/

يمكنك أيضاً متابعة الدكتور نزار حبش والدكتور مصطفى جرار لمعرفة أهم التحديات التي تواجه معالجة اللغة العربية والتوجهات الحالية، مثلاً:

https://fb.watch/a2xkE63XAB/

https://fb.watch/a2xoKiA_HL/

https://wti.kaust.edu.sa/upcoming-events/Machine-Learning-Arabic-NLP-Webinar

أين يمكن أن أجد dataset؟

يوجد الكثير من المواقع التي تتيح لك بيانات مجاناً أهمها وبدون ترتيب محدد في مجال معالجة اللغات الطبيعية:

https://nlpprogress.com/

https://machinelearningmastery.com/datasets-natural-language-processing/

https://en.wikipedia.org/wiki/List_of_datasets_for_machine_learning_research#Text_data

https://metatext.io/datasets-list/arabic-language

https://arbml.github.io/masader/

https://linguistics.stanford.edu/resources/resources-corpora

https://universaldependencies.org/

https://nyuad.nyu.edu/en/research/faculty-labs-and-projects/computational-approaches-to-modeling-language-lab/resources.html

في مجالات الذكاء الصنعي بشكل عام:

https://www.kaggle.com/datasets

https://huggingface.co/datasets

https://datahub.io/collections

https://www.paperswithcode.com/datasets

https://gluebenchmark.com/tasks

http://research.microsoft.com/en-US/projects/data-science-initiative/datasets.aspx

https://ai.facebook.com/datasets/

https://data.world/

https://www.data.gov/

https://dev.socrata.com/data/

https://www.reddit.com/r/datasets/

https://archive.ics.uci.edu/ml/index.php

https://registry.opendata.aws/

https://data.worldbank.org/

https://datasetsearch.research.google.com/
https://visualdata.io/discovery
https://developer.ibm.com/exchanges/data/
http://snap.stanford.edu/data

يمكنك القيام بجمع الداتا الخاصة بك وهي بحد ذاتها قيمة مضافة في كثير من الأحيان، ولكنها عملية مكلفة جداً وخصوصاً لو احتجت إلى تنميطها، كما أنها ليست بالضرورة أن تكون جيدة لحل المشكلة، تأكد عند اتخاذ قرار جمع الداتا الخاصة بك أن تمتلك الوقت الكافي لذلك بحيث لا تقضي كل الوقت المتاح للعمل على المشروع في جمع الداتا فقط.

كيف أقوم بصياغة المسألة؟

بعد تحديد الفكرة والقيام بالدراسة المرجعية من المهم جداً صياغة المسألة بطريقة رياضية، واضحة، ودقيقة بحيث تبقى هذه الصياغة هي البوصلة التي تجنبك الانحراف عن المشكلة والهدف، وفي هذه الصياغة يجب أن تحدد ما يلي:

1. تحديد المشكلة والوضع الحالي initial state بدقة، على سبيل المثال: المشكلة هي التنبؤ بأسعار المنازل.
2. تحديد قيود وحدود المسألة، على سبيل المثال:
   - وجود عدد محدود جداً من عينات البيانات.
   - دقة الكاميرات المستخدمة في أنظمة المراقبة.
3. تحديد هدف final state دقيق وواضح وقابل للقياس والتحقيق أي يحقق ما يسمى SMART، وعليه يمكن تحديد الدخل والخرج بشكل رياضي.
4. تحديد فضاء الحلول solution space: وهو كل التقنيات والحلول المتاحة التي تمكنك من الوصول للهدف والتي ستحصل عليها من الدراسة المرجعية.
5. حدد مصادر البيانات المتاحة.
6. حلل المسألة وضع خطة واضحة للتنفيذ تحدد فيها دخل وخرج كل مرحلة والوقت اللازم لإنجازها.

كيف أقوم بالتنفيذ؟

- بعد صياغة المسألة وتحديد مجموعة البيانات التي ستعمل عليها، قم بتقسيم هذه البيانات إلى أقسام: training dataset، validation dataset, testing dataset، مع العلم أن هذا التقسيم قد يتم بشكل مختلف احياناً و بتسميات مختلفة وحتى بأقسام أكثر، ومن المثير للاهتمام أن نذكر هنا أن بعض الجامعات تطلب التقسيم إلى أربع مجموعات: المجموعات الثلاث السابقة مع مجموعة تدعى dev dataset، بحيث يجري الاختبار عند تجريب أكثر من نموذج أو حل على هذه المجموعة بدلاً من testing dataset بل وإضافة إلى ذلك يُمنع استخدام مجموعة الاختبار أكثر من عدد محدود من المرات عند التنفيذ (على سبيل المثال 3 مرات على الأكثر)! بحيث يجري تقييم أفضل نموذج حصلت عليه تقييماً تهائياً فقط على مجموعة الاختبار، السبب في ذلك أن تحسين النموذج بالاعتماد على نتائج

أدائه على مجموعة الاختبار يعد نوعاً من الغش وتسريب للبيانات data leakage، حيث يجري تحسين هذا النموذج لكي يكون جيداً على مجموعة الاختبار وكأنك تحاول إجراء fit عليها بشكل غير مباشر ليتناسب معها ومع الأنماط الموجودة فيها.

- بعد تقسيم البيانات قم بتحليلها (باستثناء مجموعة الاختبار) وعرضها لفهمها واكتشاف الأخطاء الموجودة فيها والتي سوف تؤثر إن لم تجر معالجتها على النتائج النهائية بشكل مرعب، تذكر أن:

Garbage in -> Garbage out

- قم بمعالجة البيانات وتصحيح الأخطاء الموجودة فيها، ولا تنس أن عمليات تقسيم البيانات يجب أن تسبق عملية المعالجة لكي تمنع حدوث التسريب، بحيث تعالج بيانات التدريب اولاً، وعند القيام بتقييم أداء النموذج تطبق على مجموعة الاختبار نفس العمليات التي طبقتها على مجموعة التدريب دون معرفة أي شيء عنها، مثال:

عندما نعالج القيم المفقودة في مجموعة التدريب وندرب النموذج وننتقل لنختبره, فأول ما نقوم به لإختبار النموذج هو أن نعالج مجموعة الاختبار بنفس طريقة علاج التدريب, فلو مثلاً أردنا تعبئة القيم المفقودة بالمتوسط الحسابي، نقوم بحسابه من مجموعة التدريب فقط، ونحتفظ به ونستخدمه لمعالجة القيم المفقودة في مجموعة الاختبار دون الحاجة لحساب وسط حسابي جديد لها.

- تذكر تغير ترتيب مجموعة البيانات بشكل عشوائي shuffling حتى لا يقوم النموذج باكتشاف أنماط خاطئة لها علاقة بالترتيب.
- عرف مقياس أو مقاييس التقييم evaluation metrics التي تمكنك من:
  1. مقارنة حلك مع الأعمال السابقة التي عملت على نفس المهمة وعلى نفس مجموعة البيانات.
  2. مقارنة النماذج الخاصة بك مع النموذج البدائي baseline.
  3. قياس تحسن النموذج.

ويمكنك معرفة المقياس المناسب لمسألتك ايضاً من الدراسة المرجعية، وبدونه لن يكون هدفك قابل للقياس.

يجب اختيار هذا المعيار بحيث يكون رقمي وغير مبهم الدلالة.

- عرف نموذج بدائي baseline قوي لتقارن أداء حلولك مع أداءه، ويجب ألا يكون بسيط أكثر من اللازم حتى تستطيع تحديد فيما إذا كان تعقيد النموذج يضيف قيمة للعمل.
- من المعروف أن نماذج تعلم الآلة وعلى وجه الخصوص الشبكات العصبونية يصعب فيها اكتشاف الأخطاء debugging، وتزداد صعوبة الأمر كل ما ازداد تعقيد النموذج، من أجل تسهيل هذه المهمة قدر الإمكان لا تبدأ العمل على كل مجموعة التدريب بل خذ قسم صغير جداً منها للتأكد فيما إذا كان حلك يعمل وينتج شيء معقول، بعد التأكد بإمكانك العمل على كل المجموعة.
- لا تنس ضبط كل المعاملات غير القابلة للتعلم hyperparameters tuning.
- الشبكات العصبونية دوماً جائعة للتعلم إن صح التعبير, وما من شيء أسهل عليها من إيجاد shortcuts للأنماط التي يجب اكتشافها، بإمكانك الاطلاع على الحوادث الناجمة عن هذه القدرة لدى الشبكات العصبونية في الرابط أدناه، وهو عبارة عن database لكل الأخطاء والحوادث التي أنتجها الذكاء الصنعي على مر السنين:

https://incidentdatabase.ai/summaries/incidents

إذاً في حال لم تتعلم الشبكة فإنه يوجد خطأ لديك، إما في الكود أو في بناء النموذج نفسه.

"Neural networks want to learn!"

- من الجيد بعد الانتهاء من التجارب والتدريب والحل القيام ببعض التحليل للأخطاء التي أنتجها الحل Error analysis، على سبيل المثال محاولة تفسير لماذا أخطأ النموذج في تصنيف صنف معين دون غيره في مسائل التصنيف الآلي، وما هي الأصناف التي خلط بينها وبين هذا الصنف، هل يمكن حل هذا الخطأ بإضافة المزيد من عينات التدريب، هل عينات التدريب منحازة، ... الخ، وحتى بإمكانك بعد هذا التحليل لتأييد صحة تفاسيرك في بعض الحالات تقسيم معيار التقييم حسب إنتاج الأخطاء، مثلاً لو وجدت في مسائل الترجمة الآلية أن النماذج تخطئ أكثر على الجمل الأطول، قم بحساب معيار تقييم النموذج على العينات حسب طولها (مخطط علاقة الدقة مثلا بأطوال الجمل).
- ناقش إمكانية إضافة تحسينات على العمل لاحقاً، وهل هي متاحة حالياً.

الأمر ليس سهلاً كما يبدو، يمكنك الاطلاع على هذه الأعمال لفهم صعوبات وتحديات حلول الذكاء الصنعي:

Why AI is Harder Than We Think: https://arxiv.org/pdf/2104.12871.pdf

How to avoid machine learning pitfalls: a guide for academic researchers: https://arxiv.org/pdf/2108.02497v1.pdf

Artificial Intelligence is stupid and causal reasoning won't fix it: https://arxiv.org/pdf/2008.07371.pdf

The Bitter Lesson: http://www.incompleteideas.net/IncIdeas/BitterLesson.html

طبعاً الآراء الواردة تعكس وجهات نظر أصحابها وتجاربهم وليس بالضرورة أن تكون جميعها صحيحة.

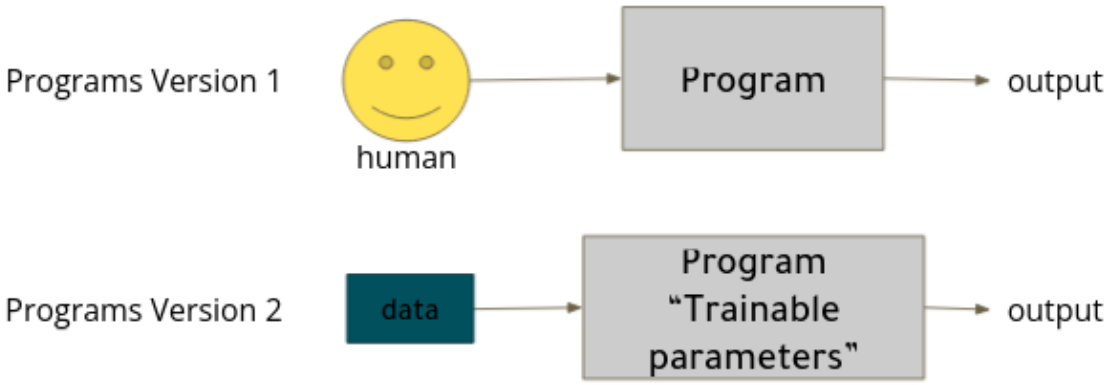
ماذا لو كان حل المسألة تحليلي analytical solution ولا يحتاج إلى تعلم آلي؟

يخطئ الكثير وخصوصاً المبتدئين في هذا المجال باعتبار أن التعلم الآلي هو الذكاء الصنعي والذكاء الصنعي هو التعلم الآلي وهذا غير صحيح! عند وجود حل تحليلي أو بعبارة أخرى عندما نعرف تابع الربط بين الدخل والخرج من غير المفيد أن تستخدم التعلم الآلي، ومن جهة أخرى من الخطأ أيضاً الاعتقاد أن حل المسألة مثلاً عن طريق القواعد لا يعني تقسيم البيانات إلى مجموعات منفصلة للتدريب والاختبار، رغم عدم وجود مفهوم التدريب هنا إلا أن تطوير القواعد الخاصة بك يحتاج لتعديل (على سبيل المثال كتابة قواعد بدائية ثم تجريبها على مجموعة التدريب من أجل تحديد مواطن الخلل فيها وتعديلها وتكرار هذه العملية حتى الوصول إلى قواعد جيدة في حال عدم وجود خبير لكتابة القواعد عنك، وكذلك الأمر عندما تطور حل تحليلي خاص بك) وفي النهاية الاختبار على مجموعة الاختبار لقياس أداء النظام، إذاً عدم وجود تعلم آلي لا يعني عدم وجود اختبارات على مجموعة منفصلة من البيانات.
تذكر أن الاختبار والتقييم وتحليل الأخطاء من أهم مراحل عملك بعد مرحلة الدراسة المرجعية.

لماذا تفشل 85% من مشاريع الذكاء الصنعي؟

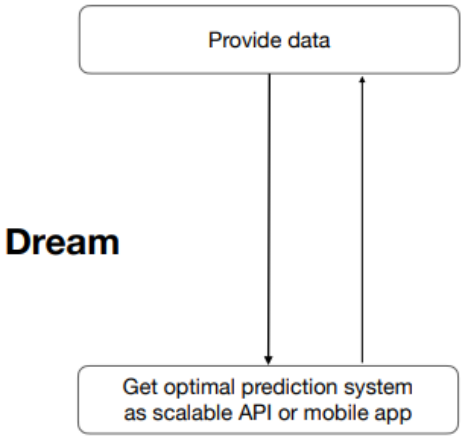
يوجد الكثير من الأسباب وراء فشل معظم مشاريع الذكاء الصنعي، وهو أمر محبط قليلاً لكن معرفة والاعتراف بالمشكلة هو نصف الحل، سنعرض تالياً بعض وأهم هذه الأسباب:

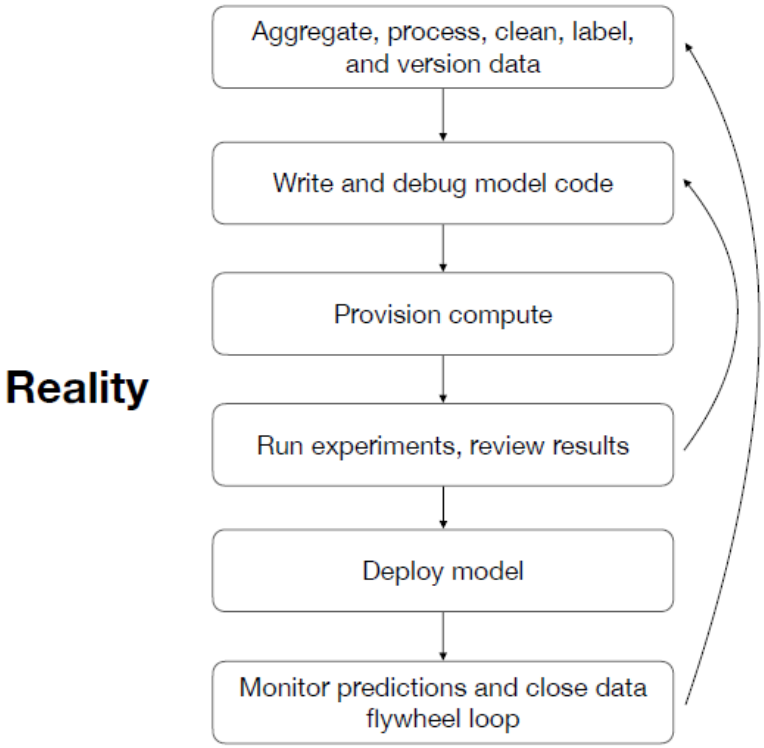
1. بناء الآمال والتوقعات العريضة من جراء إدخال الذكاء الصنعي على الأعمال businesses دون دراسة كافية وفهم هذا العمل business الذي نريد أو يريد إدخال نفسه في ورطة الذكاء الصنعي، ودون وضع استراتيجية جيدة يُحدد فيها الهدف بدقة لضمان الوصول إليه بحيث تتلاءم مع استراتيجية عمل المؤسسة أو الشركة، لمجرد الانجرار وراء تيار الذكاء الصنعي وركوب الموجة، مثلاً رغبة الشركة بأن تكون السباقة في استخدام مجيب آلي لخدماتها، أو الرغبة باستخدام تقنية ما لمجرد أن شركة كبرى في هذا المجال مثل غوغل أعلنت عنها، وبالتالي يتحول الذكاء الصنعي إلى هدف بعينه بدلاً من أن يكون حلاً لمشكلة أو أداة لتحقيق الهدف، ويتحول فيه هذا الهدف إلى مشكلة فيما بعد مع الانجرار كل مرة وراء أحدث التقنيات.
2. كما نعلم أن الذكاء الصنعي مقاد بالبيانات، بدون البيانات لا يوجد حل ذكي:

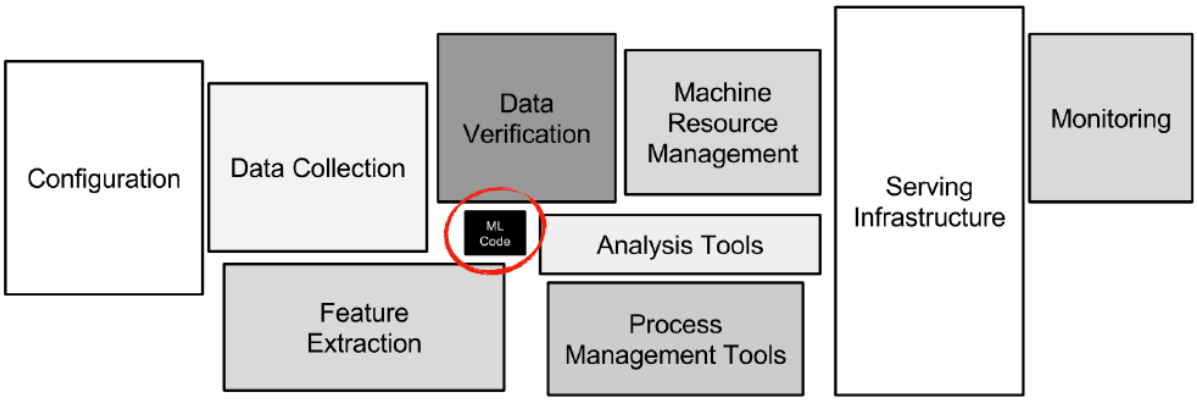


وعليه يجب أن نسأل أنفسنا قبل البدء هل يوجد ما يكفي من البيانات للعمل؟ هل تمتلك المؤسسة هذه البيانات؟ ما هو حجم البيانات الكافي؟ إجابة هذه الأسئلة تعتمد على حالات الاستخدام use cases والنتائج المتوقعة ونوع البيانات.

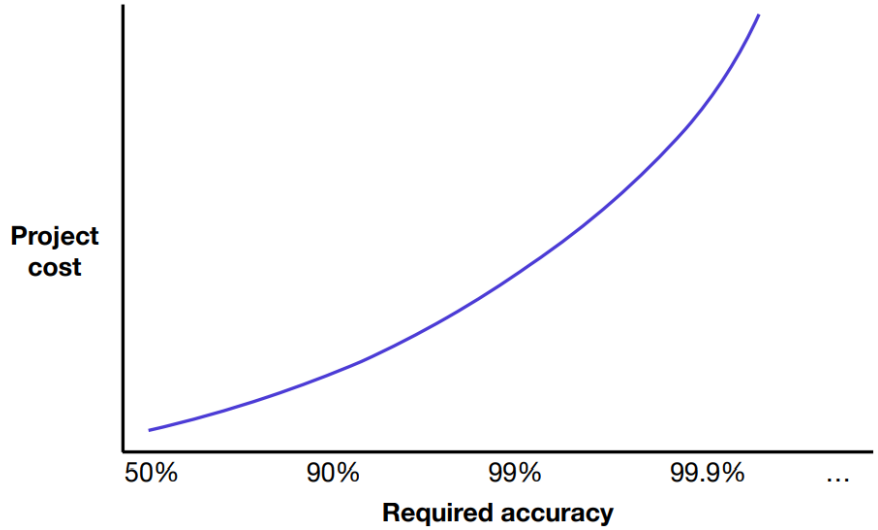
3. على عكس المتوقع، من الأسباب التي تقود للفشل هو قلة الخبراء في هذا المجال! نعم، لا يوجد الكثير من الخبراء خارج منطقة (داتا-نموذج-تدريب) في باقي مراحل مشروع الذكاء الصنعي مثل MLops و DataOps و Data engineering... الخ، وكذلك في تهيئة البيئة ومعرفة البنية التحتية المناسبة لعمل المشروع، لاحظ في المخططات التالية الفرق بين الحقيقة والتوقع:


Provide data
Dream
Get optimal prediction system
as scalable API or mobile app


Aggregate, process, clean, label,
and version data
Write and debug model code
Provision compute
Reality
Run experiments, review results
Deploy model
Monitor predictions and close data
flywheel loop



4. نقص الخبرة والتعاون بين أعضاء فريق الذكاء الصنعي، ونقص التعاون أيضاً مع باقي فرق المؤسسة والعمل بشكل منعزل. بدون هذا التعاون تبدأ أولى مرحلة من مراحل الفشل.
5. قلة المعرفة بالمجال أيضاً لدى العامة المأخوذون بفكرة استبدال الآلة بالبشر مما يقلل من حجم الاستثمارات في هذا المجال الذي هو بالأساس مكلف جداً (كلفة المعطيات وال gpu والبنية التحتية ... الخ)، لاحظ علاقة دقة النموذج بالكلفة في المخطط:



6. يظن أعضاء فريق الذكاء الصنعي أنه لا يحتاج إلى قواعد هندسة البرمجيات للعمل، وتبدأ كتابة ال spaghetti code مع العلم أن نماذج الذكاء الصنعي أساساً صعبة الصيانة واكتشاف الأخطاء كما ذكرنا سابقاً فكيف هو الأمر مع نقص قواعد الهندسة!
7. صعوبة مكاملة الحل الذكي مع النظام الموجود بالفعل.
8. تحويل أي مسألة مهما كانت بسيطة إلى مسألة تعلم آلة ولو كان حلها بقواعد المعرفة يعطي نتائج جيدة جداً (كلف إضافية بدون جدوى).
9. القفز مباشرة إلى النماذج المعقدة والعمل دفعة واحدة بدلاً من العمل بشكل تكراري.

نصائح أخرى

تابع هذا الكورس الرائع المقدم من berkeley عن كيفية تحضير وإدارة مشروع تعلم عميق من الصفر:
https://fullstackdeeplearning.com/

يمكنك استخدام هذه العبارات عند البحث عن أكواد لمشروع ما لمعرفة كيفية كتابة كود جيد والأخطاء التي يمكن أن تحدث في المهمة التي تعمل عليها:

- [task_name] Best practices
- [task_name] Tips and tricks
- Awesome [task_name]

من الجيد الاطلاع على أعمال الغير حتى من ناحية الكود (الاطلاع وليس السرقة).

المصادر:

http://web.stanford.edu/class/cs224n/readings/final-project-practical-tips.pdf
https://www.youtube.com/watch?v=733m6qBH-jI&t=3s
https://web.mit.edu/tslvr/www/lessons_two_years.html
https://writingcenter.unc.edu/tips-and-tools/literature-reviews/
http://www.aspu.edu.sy/laravel-filemanager/files/18/3%20-%20%D8%AF%D8%B1%D8%A7%D8%B3%D8%A9%20%D9%85%D8%B1%D8%AC%D8%B9%D9%8A%D8%A9.pdf
https://developers.google.com/machine-learning/problem-framing/formulate
https://medium.com/@NeotericEU/the-single-most-important-reason-why-ai-projects-fail-68c8c1f72dee#:~:text=Research%20shows%20that%20there%20are,you%20actually%20start%20your%20project.
https://towardsdatascience.com/why-85-of-ai-projects-fail-2c85cdd4bf45

ملاحظة هامة:

لا تقم بكتابة المراجع أو توثيق عملك بنفس الطريقة المتبعة هنا!!