



الجمهورية العربية السورية
المعهد العالي للعلوم التطبيقية والتكنولوجيا
قسم المعلومات
العام الدراسي 2017/2018

تقييم جودة نصوص اللغة الانكليزية

مشروع السنة الرابعة

إعداد

فاروق حجابو

إشراف

م. رياض سنبل

د. غيداء ريداوي

2 آب 2018

الملخص

المحتويات

i	الغلاف
ii	الملخص
iii	المحتويات
v	قائمة الأشكال
vi	قائمة الجداول
vii	الاختصارات
1	1 التعريف بالمشروع
1	1.1 مقدمة
2	2 الدراسة المرجعية
2	1.2 تعلم الآلة
3	1.1.2 تصنيفات تعلم الآلة

4 خوارزميات تعلم الآلة 2.1.2

5 المراجع

قائمة الأشكال

قائمة الجداول

الاختصارات

AI	Artificial Intelligence
ML	Machine Learning
NLP	Natural Language Processing
SVM	Support Vector Machines

الفصل الأول

التعريف بالمشروع

يُمهّد هذا الفصل للمشروع، حيث يُبيّن فكرة المشروع وأهميتها والأهداف المرجّوة منه. ويذكر المتطلبات الوظيفية وغير الوظيفية للمشروع.

1.1 مقدمة

الفصل الثاني

الدراسة المرجعية

يبيّن هذا الفصل الدراسة المرجعية للمشروع. يبدأ بتقديم مفاهيم تعلّم الآلة ومراحلها المختلفة والمعايير المعتمدة لتقييمها. ويقدم مفاهيم ومراحل معالجة اللغات الطبيعية. وأخيراً يسرد بعض الأوراق الأبحاث العلمية المتعلقة بالمشروع، ويوضح المنهجيات المتبعة فيها.

1.2 تعلم الآلة

تعلم الآلة Machine Learning هو فرع جزئي من الذكاء الصناعي Artificial Intelligence. يُقصد بتعلم الآلة مجموعة الأدوات والمفاهيم والمنهجيات المستخدمة لبرمجة الحواسيب بطريقة تسمح لهذه الحواسيب بالتعلم من المعطيات [11].

ويمكن أيضاً تعريفه بشكل أكثر عمومية كالتالي:

“Machine Learning is the field of study that gives computers the ability to learn without being explicitly programmed.”

—Arthur Samuel, 1959

كما يعتبر التعريف التالي تقني وأكثر دقة:

“A computer program is said to learn from experience E with respect to some task T and some performance measure P , if its performance on T , as measured by P , improves with experience E .”

–Tom Mitchell, 1997

على سبيل المثال، النظام الذي يقوم بفلتر الإيميلات إلى إيميلات مؤذية spam وإيميلات سليمة، يستخدم منهجيات تعلم الآلة. يقوم هذا النظام بتعلم طريقة التمييز بين هذين النوعين من الإيميلات باستخدام عدد كبير من الأمثلة والمعطيات المصنفة مسبقاً. نسمي هذه المجموعة من الأمثلة بمعطيات التدريب training set، وكل مثال منها نسميه مثال تدريبي training instance.

في هذه الحالة، المهمة T هي تصنيف الإيميلات الجديدة إلى إيميلات مؤذية وغير مؤذية، الخبرة E هي مجموعة معطيات التدريب، ومؤشر قياس الأداء P يمكن تعريفه بعدة طرق؛ فمثلاً يمكننا استخدام نسبة نسبة عدد الإيميلات التي تم تصنيفها بشكل صحيح إلى عدد الإيميلات الكلي.

1.1.2 تصنيفات تعلم الآلة

يمكن تصنيف أنظمة تعلم الآلة وفق عدة معايير. التصنيف الأكثر شهرة يعتمد على آلية التدريب، وهو كالتالي:

- التعلم تحت الإشراف supervised learning: وهي حالة أن تكون الأمثلة التدريبية متوفرة مع الخرج label المرتبط بها. وهذه حالة مثال تصنيف الإيميلات المطروح سابقاً. حيث أن معطيات التدريب هي مجموعة كبيرة من الإيميلات المصنفة مسبقاً من قبل البشر إلى إيميلات مؤذية وغير مؤذية.
- التعلم بدون إشراف unsupervised learning: وهي حالة أن تكون معطيات التدريب موجودة ولكنها غير مصنفة unlabeled أو غير مرتبطة بخرج معين. على سبيل المثال، قد ترغب شركة في تصنيف زبائنهم إلى عدة مستويات، زبائن من الدرجة الأولى، زبائن من الدرجة الثانية، وهكذا. فيمكن استخدام تعلم الآلة لاكتشاف بعض الأنماط الموجودة في معطيات الزبائن واكتشاف هكذا تصنيف. وهذا ما يُعرف بالتجميع clustering.
- التعلم نصف المشرف عليه semi-supervised learning: وهي حالة وسيطة بين التصنيفين السابقين. تكون فيها بعض أمثلة التدريب مرتبطة بخرج معين (غالباً تشكل النسبة صغيرة)، وتكون باقي الأمثلة غير مرتبطة بخرج. تنطبق هذه الحالة على مثال تصنيف الإيميلات في حال لم تكن جميع معطيات التدريب مصنفة بشكل مسبق.
- التعلم بالتعزيز reinforcement learning: وهي الحالة التي يتخاطب فيها النظام مع بيئة أخرى. تقدم له هذه البيئة نتائج feedback بناءً على أفعاله. هذا الصنف ينطبق على الخوارزميات المستخدمة لتدريب الأنظمة

التي تتعلم الألعاب. حيث يقوم النظام بمجموعة من الأفعال actions ضمن بيئة اللعبة، وبناءً على النتائج (تحسن نتيجته أو انخفاضها) يغير أفعاله اللاحقة.

وعلى وجه الخصوص يمكن تصنيف التعلم تحت الإشراف supervised learning بحسب نوع الخرج المرتبط بمعطيات التدريب. سنصفها بشكل أساسي عريض كالتالي:

- التصنيف classification:

- regression

2.1.2 خوارزميات تعلم الآلة

English References

- [1] Sarah E Schwarm and Mari Ostendorf. “Reading level assessment using support vector machines and statistical language models”. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics. .2005
- [2] Rong Zheng et al. “A framework for authorship identification of on-line messages: Writing-style features and classification techniques”. In: Journal of the American society for information science and technology 57.3 .(2006)
- [3] Lijun Feng, Noémie Elhadad, and Matt Huenerfauth. “Cognitively motivated features for readability assessment”. In: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics. .2009
- [4] Sarah E Petersen and Mari Ostendorf. “A machine learning approach to reading level assessment”. In: Computer speech & language 23.1 .(2009)
- [5] Xiaofei Lu. “Automatic analysis of syntactic complexity in second language writing”. In: International journal of corpus linguistics 15.4 .(2010)
- [6] Sowmya Vajjala and Detmar Meurers. “On improving the accuracy of readability classification using insights from second language acquisition”. In: Proceedings of the seventh workshop on building educational applications using NLP. Association for Computational Linguistics. .2012

- [7] Sowmya Vajjala and Detmar Meurers. “Readability assessment for text simplification: From analysing documents to identifying sentential simplifications”. In: IJL-International Journal of Applied Linguistics 165.2 .(2014)
- [8] Sowmya Vajjala. “Analyzing text complexity and text simplification: connecting linguistics, processing and educational applications”. PhD thesis. Ph. D. thesis, University of Tübingen, .2015
- [9] Sowmya Vajjala and Detmar Meurers. “Readability-based sentence ranking for evaluating text simplification”. In: arXiv preprint arXiv:1603.06009 .(2016)
- [10] Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. “Text readability assessment for second language learners”. In: Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications. .2016
- [11] Aurélien Géron. Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems. ” O’Reilly Media, Inc.”, .2017
- [12] Sowmya Vajjala and Ivana Lucic. “OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification”. In: .(2018)

المراجع العربية

- [13] رياض سنبل. استخلاص مخططات إجراءات العمل انطلاقاً من النصوص Extracting Business Process Models from Natural Language Texts .2017. المعهد العالي للعلوم التطبيقية والتكنولوجيا.