# Projet MA1b
# Apprentissage incrémental de plongements lexicaux

Alexandre Kabbach
alexandre.kabbach@unige.ch

Paola Merlo
paola.merlo@unige.ch

20.02.2020

## 1 Introduction

Les plongements lexicaux [Bengio et al.2003, Collobert et al.2011, Huang et al.2012, Mikolov et al.2013, aka *word embeddings*] et leurs implémentations basées sur des réseaux de neurones telles que `Word2Vec` [Mikolov et al.2013, ci-après W2V] sont particulièrement sensibles à la taille du corpus d'entraînement.

Récemment, plusieurs travaux ont proposé d'opérer des modifications à l'architecture de W2V pour permettre l'aquisition de plongements lexicaux de manière *incrémentale* et sur la base de *faibles quantités* de données [Herbelot and Baroni2017, Kabbach et al.2019].

L'objectif de ce projet est d'intégrer la proposition de [Kabbach et al.2019] à l'architecture W2V sous Tensorflow, en remplaçant le filtre de *subsampling* traditionnel de W2V par un filtre basé sur leur notion d'*informativité du contexte*, puis d'étudier l'impact de ce filtre sur la qualité et la vitesse d'acquisition des plongements lexicaux.

Pour ce faire, nous mettons à votre disposition une version pré-implémentée de W2V sous Tensorflow, disponible ici : `https://github.com/akb89/word2vec`.

## 2 Détails et planning

### 2.1 Prise en main

Les deux premières semaines du projet seront consacrées à la prise en main de W2V. Pour se faire, vous réaliserez le TP4 du cours METL. Date de soumission : 8 mars 2020.

### 2.2 Formalisation

Dans un second temps, vous vous baserez sur l'article [Kabbach et al.2019] et rédigerez un court plan d'implémentation détaillant votre méthologie pour intégrer la notion d'informativité au subsampling de W2V. Ce plan détaillera les parties de code que vous comptez modifier, vos motivations pour le faire, et vos besoins supplémentaires en matière d'implémentation (par exemple, si vous comptez développer des fonctionalités supplémentaires dans Tensorflow). Date de soumission : 15.03.2020.

### 2.3 Implémentation

Vous vous attaquerez ensuite à l'implémentation à proprement parler. Vous veillerez à déposer votre code sur Github en le versionnant régulièrement. Date de soumission : 05.04.2020.

### 2.4 Expérimentation

Vous procéderez ensuite à une série d'expériences en faisant varier les hyperparamètres détaillés dans [Kabbach et al.2019] et en comparant les performances du modèle W2V modifié à celles du modèle standard. Dans un premier temps, vous limiterez l'évaluation su modèle à sa correlation avec un jeu de données de similarité lexicale [Bruni et al.2014, MEN]. Dans un second temps vous proposerez une méthode d'évaluation permettant de comparer les vitesses de convergences des différents modèles. Date de soumission : le 08.05.2020 (avec le rapport final).

## 3 Rendus

1. TP4 : à soumettre par email avant le 8 mars 2020
2. Rapport d'implémentation : à soumettre par email avant le 15.03.2020
3. Implémentation : à soumettre sur Github avant le 05.04.2020
4. Rapport final : à envoyer par email avant le 08.05.2020
5. Présentation orale : prévue la semaine du 11.05.2020

Ce projet prévoit une centaine d'heures de travail. La note finale sera basée sur la version finale du code, du rapport et de la présentation orale.

## References


[Bengio et al.2003] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155, March.

[Bruni et al.2014] Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 49:1–47.

[Collobert et al.2011] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12:2493–2537, November.

[Herbelot and Baroni2017] Aurélie Herbelot and Marco Baroni. 2017. High-risk learning: acquiring new word vectors from tiny data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 304–309, Copenhagen, Denmark, September. Association for Computational Linguistics.

[Huang et al.2012] Eric Huang, Richard Socher, Christopher Manning, and Andrew Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 873–882, Jeju Island, Korea, July. Association for Computational Linguistics.

[Kabbach et al.2019] Alexandre Kabbach, Kristina Gulordava, and Aurélie Herbelot. 2019. Towards incremental learning of word embeddings using context informativeness. In *Proceedings of the 57th Conference of the Association for Computational Linguistics: Student Research Workshop*, pages 162–168, Florence, Italy, July. Association for Computational Linguistics.

[Mikolov et al.2013] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3111–3119, USA. Curran Associates Inc.