Analyse statistique et empirique des modèles de Word Embedding sur Twitter

Kim Antunez, Romain Lesauvage, Alain Quartier-la-Tente

sous l'encadrement de Benjamin Muller (doctorant à l'Inria)

1 Contexte

Grâce à l'évolution des méthodes d'apprentissage profond (*Deep Learning*), l'appréhension du langage naturel est aujourd'hui devenu une discipline à part entière ("*Natural Language Processing*). Ce succès s'explique en partie grâce à l'émergence de techniques non supervisée d'apprentissage de représentation de structures linguistiques. Les méthodes de *word embedding* (« plongement lexical » en français) permettent de représenter chaque mot d'un dictionnaire par un vecteur de nombres réels afin que les mots qui apparaissant dans des contextes similaires possèdent des vecteurs correspondants qui sont relativement proches. Les modèles **word2vec**, développés par une équipe de recherche chez Google sous la direction de Tomas Mikolov, sont parmi les plus célèbres.

2 Objectif du projet

Dans ce projet de statistiques appliquées, nous étudierons dans un premier temps en détail et implémenterons le modèle *word2vec*. Puis, dans un second temps, nous mettrons ce modèle en application sur une base de plusieurs dizaines de millions de tweets (2014-2017).

3 Travail effectué

3.1 Compréhension du modèle

Nous avons débuté le projet en nous imprégnant du champ des méthodes de NLP et même de Deep-learning qui nous était jusqu'alors inconnu. Pour cela nous avons lu les articles présentés dans la bibliographies ainsi que d'autres articles de blogs, en particulier concernant l'implémentation sur python

3.2 Implémentation du modèle

Nous avons tous les trois implémenté le modèle individuellement sur python en utilisant la librairie Pytorch et avons testé

3.3 Evaluation du modèle implémenté

Malgré leurs utilisations presque généralisées, très peu de travaux théoriques expliquent ce qui est réellement capturé par ces représentations de mots. C'est pourquoi l'évaluation de l'efficacité de ce modèle ne peut se faire qu'à l'aide de méthodes empiriques.

Nous avons commencé à évaluer qualitativement le modèle en l'entraînant sur des données fictives.

évaluation quali

4 Perspectives

4.1 implémentation et évaluation

- mise en commun des codes des 3 membres du groupes et implémentation d'un modèle "propre"
- faire tourner le modèle sur les vraies données et si besoin tenter d'optimiser son implémentation en repérant les parties du codes qui mettent du temps à tourner.
- nearest neighbor , human judgment agreement

4.2 analyse

Tout le travail d'analyse à partir de la base de données exhaustive des tweets reste à effectuer. Les possibilités d'applications sont nombreuses mais nous devrions a priori nous orienter sur une comparaison entre l'évolution du solde d'opinion des ménages et l'évolution de "l'humeur" des tweets (données trimestrielles de 2014 à 2017)