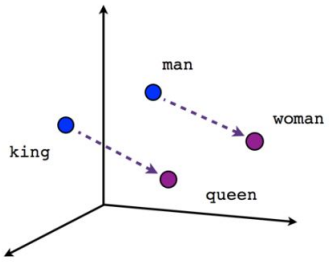
INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES TOULOUSE

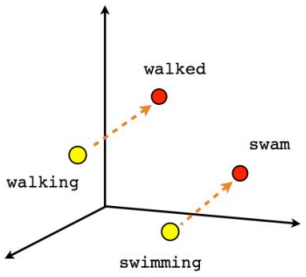
INSTITUT de MATHEMATIQUES de TOULOUSE

Projet Recherche-Innovation de 4ème année GMM 2019/2020

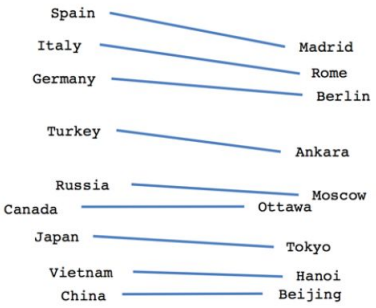
Classification non- supervisé de texte et Word Embedding.



Male-Female



Verb tense



Country-Capital

Objectif

Le traitement automatique du langage naturel a connu des avancées majeures grâce à l'intelligence artificielle ces dernières années. Traduction automatique (Google translate, DeepL), analyse de sentiment (commentaires facebook/amazon), classification (détection de spam, etc..).

Ces avancées sont notamment dues à l'apparition d'algorithme de **words-embedding**. Ces algorithmes permettent d'encoder un mot de façon "intelligente" et ce de façon automatique en parcourant des centaines de millions de lignes de textes. Cet encodage permet ainsi d'effectuer des opérations sur les mots du type : 'roi' - 'homme' + 'femme' = 'reine' ou encore 'paris' - 'france' + 'espagne' = 'madrid' (voir illustration ci-dessus) et ainsi solutionner des problèmes complexes cités plus haut.

Dans ce projet, nous appliquerons ces modèles sur des données textuelles classiques (wikipedia) et des données textuelles aux propriétés particulières : descriptions de produit Cdiscount. L'objectif est d'étudier la performance de ces modèles sur ces types de données et d'étudier leurs propriétés. Enfin nous essayerons d'adresser des problèmes de NLP telle que la classification de texte ou l'extraction/résumé d'informations à l'aide des représentations générées par ces modèles.

Travail demandé

Le premier travail consistera à bien s'approprier les algorithmes de word-embedding les plus utilisés (Word2Vec[1], Glove [3] ou FastText[4]). Ensuite on appliquera ces méthodes sur les descriptions de produits de Cdiscount mis à disposition à l'aide de librairie python. Un travail d'exploration (Clustering, ACP, TSNE). sera demandé sur les représentations des descriptions de produits générées par ces modèles. Ensuite on utilisera les modèles entraînés sur le jeu de données d'apprentissage afin d'adresser les problèmes NLP décrits précédemment. Tout ce travail sera réalisé en python.

Référence

[1] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781. •

[2] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111-3119). • [3] Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).

[4] Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, 5, 135-146.

Tuteur : Brendan Guillouet