TD6 : Topic Modeling

LO17

Fatma Chamekh

2021-2022

Objectifs :

L'objectif de ce TD est d'explorer le topic modeling en utilisant l'algorithme LDA. Il ya aussi un aperçu sur la similarité entre les documents. Le Td comporte deux parties :

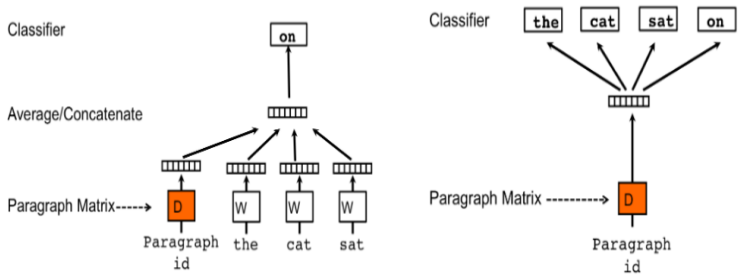
- Partie 1 : la classification doc2vec.
- Partie 2 : Découverte et manipulation du modèle LDA.

Partie 1 :

Dov2Vec

Doc2Vec est une extension des approches Word2Vec dans lesquelles on ajoute un "token" associé à chaque document (ici, un paragraphe). Il existe deux versions de cet algorithme (Le and Mikolov, 2014) :

- PV-DM : Distributed Memory Models of Paragraph Vectors
- PV-DBOW : Distributed Bag of Words version of Paragraph Vector



On reprend le corpus de l'exercice 1 de partie 2 du TD précèdent. Le corpus met à votre disposition d'autres informations qui vous permettent de rapprocher deux documents lorsque : a) les articles partagent un ou plusieurs auteurs en commun, b) un article cite un autre article dans sa bibliographie, c) les articles ont été publiés dans le même journal ou la même conférence. La similarité entre deux documents peut donc se baser sur leur similarité textuelle (le vecteur des mots généré via word2vec) mais également sur d'autres informations de proximité.

Une tâche intéressante consiste à essayer de trouver le nom des auteurs d'un article à partir de sa description textuelle. Cette tâche peut être d´définie comme un problème de recherche d'information dans laquelle on utilise un vecteur qui représente un auteur et on compare ce vecteur avec celui des documents. Une solution serait d'utiliser Doc2Vec en utilisant comme tag le nom de l'auteur, ce qui permet de calculer des représentations d'auteur.

Pour vous aider à implémenter votre solution, vous pouvez vous référer à l'exemple suivant : https://askcodez.com/doc2vec-obtenez-les-documents-les-plus-similaires.html

Partie 2 :

Exercice1 : traitement des données grand débat-transition écologique

Le grand débat national est une initiative qui a été mise en place par la présidence de la république afin de permettre aux Français de débattre/s'exprimer, d'une manière anonyme, sur des sujets clés (environnement, fiscalité …). A l'issu de cette initiative, les données ont été structurées en documents json et CSV.

Le gouvernement lance un appel auprès de plusieurs prestataires dont vous faites partie afin d'analyser les données portant sur la transition écologique. Afin de vous aider dans cette mission, un notebook avec les différentes étapes d'analyse/application du modèle LDA est mis à votre disposition afin de le comprendre et compléter.

Exercice2 : traitement des données grand débat-transition évènements

Etant satisfaite par vos résultats, le gouvernement vous sollicite une deuxième fois afin d'analyser les données qui porte sur les évènements. A vous de jouer !!!!