Du 29/09/2020 au 13/10/2020

* Choix Dataset (a priori, les films de Wikipedia)
* Documentation sur Word2vec
* Pistes : LSA (Latent Semantic Analysis) ou BERT
* Documents HTML 🡪 Preprocessing très important

Les pages web Wikipedia présentent l’avantage d’avoir une structure commune (avec l’introduction puis les différentes sections).