What has been done?

- Run full w2v on AWS, on strongest machine available → expensive
- Testing Azure Notebook → storage limitation
- Testing on my machine (PC) → ram limitation
- Made an iteration/generator methode to do line by line on dataset
- Made a wikipedia dataset splitter with balise preservation
- Pre-trained models are not retrainable, not full model → only Keyed Vectors
- Dictionary unexpected: - Lemmatization missing "library not installed first time run, "pattern"
  → Dog/NN, Do/VB, ..
- 1 CPU Core not working
- Memory allocation error on splitter also

short terme

- starting new sprint on - ANN chatbot
  - parallel algorithm
  - protocole to evaluate proactive

Get
- full w2v wikipedia
- meanwhile use smaller dataset
- Circulate and test of existing chatbot solution
- Fix memory error on splitter with a burner for the generator

Last time short terme

- Combine with ANN
- Document in report Gensim life
- Compare with: fasttext, word2vec, Glove

overall progress

- fixing the memory problem with gensim by splitting datasets by pages and use a generator by line
- Trying to use cloud based machines, but expensive and no results yet.
- Word2Vec operations working on partial model by dictionary is not as expected

Questions:

- Problematic de recentrement
- Tester les translations -> avec un set de phrases
  -> 20aine
- Case comprehision Word2Vec ->

Translation

-> perturbation Mini

-> le chat est une mamifaire
  -> Plus proche voisin
    -> vecteur appliquer au reste de la phrase
    -> vecteur de translation
  -> translation dirigé -> shift
  -> translation random

-> Word2Vec
  -> A la base pretraitement des mots
  -> Playing with word2Vec

-> Dimensions -> concepts
  -> Jump on 1 dimension
  -> Multi dimensions

-> Synonime
  -> point chat
    -> sinonime de chat
      -> quelles dimensions les plus impactés
      -> diriger dans la direction

- But comprendre l'espace Word2Vec
-> Word2Sequence

- Chapitre sur le biais

-> mettre en evidance

Doctor - man + woman = Nurse

-> Facebook, Google, -> biais dans recommendation publicité

-> poussé vers des études