What has been done?

- Ran multiple models ~10 per training
- Made sentence generator
- started report
- Played with analogies

Problems:

- Cannot do similarities with un semantically related words
  -> "Capital of Science" doesn't make sense with raw wikien
- Analogies refer to astracted words. Doctor -> "Doctor who character"
  -> often not in the right semantic space
- Pure random doesn't impact the geometrical operations
  -> Words from vocabulary are needed
- CPU Entrainement

short terme

- Plot the specific spaces
- Calculate and test of existing chatbot solution
- Compare Premade W2VC(google) vs mine
- Compare unlemmatized vs lemmatized
- playing with - ANN chatbot
  - parallel algorithm
  - protocol to evaluate proactive

overall progress

- Fixed the memory problem with gensim by splitting data sets by pages and use a generator by line
- Tryied to use cloud based machines. Expensive and not worth it for CPU computation, same perf as on heia-fr.
- Full wikipedia EN trained model : 35 et 20h on heia-fr
  -> with lemmitization
- Word2Vec operations working
- Testing on a CPU dedicated machine to train
- Generating new sentences

Todo :

Custom :

Stemming and already in vector

- Prendre un vecteur
  -> que vont dire les axes

Federer
-> dimension valeur la plus grand
-> varier uniquement cette variable
Ls si on se deplace sur l'axe est-ce
qu'on peut obtenir des choses
genre waouw ca

Ls

Rapport

Démo
-> Chatbot qui prend la main