what has been done?

- Setup of new CPU machine → ready and running
- Started Research on ANN and DNN methods to use Word Embadings
- started looking at alternative corpora data sets
- Currently building Word2Vec without lemmatization
- Playing with similarities
- Building Dictionnary is longer on CM? exacty 1h?

short terme

- playing with - ANN chatbot
  - parallel algorithm
  - protocol to evaluate proactive
- Creature and test of existing chatbot solution
- Combine with ANN
- Document in report Gensim life
- Compare with: fast text, word2Vec, Glove
- Build alternative Word2Vec with different corpora
- Play with Word2Vec
  - geometies
  - analogies (capital of science)
- Compare unlemmatized vs lemmatized
- Compare Premade W2VC google) vs mine

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: 3 5 et 20 h on heia-fr
  → with lemmitization
- Word2Vec operations working
- Testing on a CPU dedicated machine to train

Questions:

- Meeting next week? A partir mercred apres-midi

Faire :

-> - Stopword

-> Generate similar sentences

-> Use Word2Vec as it is

-> "drunk guy in a pub"

Avancer sur la doc !

Écrire tout ce qui a été fait jusqu'à présent

-> dUP

1) Versions
Dans quel mesures peut-on exploiter le W2V, sans intelligence

2) Seq2Seq, article + Implementation, avec quelque chose tout fait

-> Comparer les modules

Explorer la generation de phrases

-> pas greffer

2 semaines de redaction

-> mettre à jour le plan