what has been done?

- full wikipedia english training : 3J 20h on heia-fr
- lemmatization vs non-lemmatization
  - → without → performence → less shades
- playing with operations by vectors
  - → words
  - → still working on sentences
- tested analogies operations : Athens is to → Greece
  as Baghdad it to → Iraq

→ translation random
→ Plus proche voisin → le chat est un mamifaire
  → vecteur appliquer au reste de la phrase
  → vecteur de translation

short terme

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

- Literature and test of existing chatbot solution

- Combine with ANN
- Document in report Gensim life

- Compare with : fastText, word2Vec, Glove

→ translation dirigé → shift

overall progress

- fixed the memory problem with gensim by splitting datasets 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 : 3J et 20h on heia-fr
  → with lemmitization

- Word2Vec operations working

Prendre note des examples

->

est-un

sur base d'exemple

Exemple 10ème 20ème de Capital

-> Average of dimension

-> Vecteur de difference

-> Capital [ ]

-> Capital of science

-> Est le marié de..

-> Est-ce que les relations W2V sont les mêmes que les relations des entités

-> la force des similarités

-> Cosine (direction)

-> Normalizé

-> Non normalisé

-> Articles avec des trucs Bizzare qu'on a vu

-> why people are not using full lemized vocab

-> Thinking a metric to compare via similar

-> mine vs google

methold (wv_a, wv_b, list_de_mot, top_n)

↳ ratio du nombre de mots commun

↑ jolie contribution

Idée :

-> Evolution de la femme dans le word2vec

-> avec les backups de wikipedia