Project Proposal For Text Mining

Matthew Noack 6334589 - Louis Gauthy 6188059

Project Proposal: Find all known locations visited in *20000 Leagues Under the Sea* by Jules Verne. Will plot PLACES of MAIN PERSONS (Captain Nemo, Conseil, Professor Pierre Aronnax, and Ned Land). Book takes place on earth, based on a submarine voyage around the world. MAIN PERSONS are mostly grouped together, will be tracking their movements, the PLACES they visit, and the PERSONS the MAIN PERSONS meet.

a) For text-extraction, use both BERT and NLTK to extract all the words in the book and compare their results in the following steps.

b) Pre-processing: Will use full text from Guggenheim library, shouldn't need much pre-processing as full text is available.

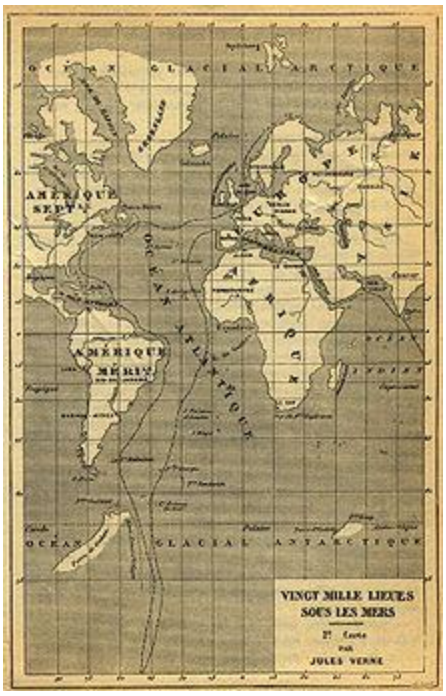
c) Basic entity extraction: Will determine PERSONS, PLACES, and other entities through creating a knowledge graph.

d) Will link PERSONS and PLACES together

e) Will detect any non MAIN PERSONS that are in the PLACES

f) Use Latent Dirichlet Allocation (LDA), and Latent Semantic Analysis (LSA) for topic modeling. Topic modeling will be used to find groups and themes within *20000 Leagues Under the Sea*.

h) To evaluate our extracted knowledge graph, we will use the follow quality criteria: - - Correctness. Will use ground truth from pathway of Nautilus from https://en.wikipedia.org/wiki/Twenty_Thousand_Leagues_Under_the_Seas.



i) We will visualize the PERSON - PLACES knowledge graphs extracted. To further accompany this static visualization, we will also visualize which PERSONS and PLACES are involved over time (of the book). Map plotting where PERSONS are in based on which PLACE they are in.