Rust: to do list. (sum up of our notes)
Don’t forget to install: openBLAS – there is a binding for rust somewhere on git

Stage 1. Preprocessing stage of preprocessing:
Collecting Corpora. (6 genres, 100 texts each, except romantic literature: but amount of texts aligned with other genres)
Stage 2. Preprocessing stage:
A) (*Using in this part*: soup)
We have texts in html (erotica) and txt (other ones) formats. Good thing that all txt texts are the same: they gonna have some kind of project gutenberg information. Probably it will make sense to remove it in preprocessing (even though tf idf will take care of it, but what if we would use it on another text corpora eventually?)
html – using type of soup to make everything equally preprocessed and txt.
B) (*Using in this part*: nltk stopwords)
we are creating two versions: one with using stopwords and one without. It is like an optional function.
Stage 2.5. (*using:* clap)
We gonna use command line, cause thanks to Peter now we know that apparently there are shit ton of pluses in doing that. *clap-clap*
Stage 3. (*using:* scanlex for word tokenizer)
-Trying to work with (uni – n)grams; here use both versions with StopWords and NoStopWords
For big texts: splitting big novels in equal parts and assuming the same label for all of them.

Algorithms: (*using:* rust-tfidf)
- Using tf/idf (can be later united with KNN) (can be applied of both types of our corpora: it would be interesting if one of “stop words” would actually appear to be significant)

(*using for KNN:* our implementation or rusty library?)
- main problem with KNN – too huge distances; so we gonna take care of that with using singular value decomposition => get a dense matrix which has fewer dimensions ergo smaller distances (*using*: ndarray)
- ndarray SVD

Also we gonna use somewhere: ordered float

So far it’s all. So basically we will experiment with several different preprocessing methods, three main algorithms etc.
If we will still gonna be energetic it would be nice to do as much more as we could to get some exceptional results, because there is a cool conference in Hong Kong about digital humanities and it would be awesome to apply.