Alok, [redacted] Deytia, Tyler, Kapil TEAM 1

Problems with data

- incorrect regex leads to missing or extra words
- vernacular words
- Extra data (title and author randomly placed in text)
- inadequate translations
- time frame from the original
-

T2

1. Issue w/ translatability - context
2. Issues of Time Period, gender - context
3. genre of works - context
4. semantics and meaning of two words with same spelling but different meanings - Context
5. Legal constraints for providing a book online for free - missing dat of

Chunyan Tang
John Duggan
R.J. Lusk

Team 3 - DJ, Mark, Sam

Issues with Data:

- Empty space as a word (stopwords + html cleaning didn't catch)
- Books not on Project Gutenberg
- Extraneous words (e.g., "Gutenberg", "copyright")
- Words that don't carry significant meeting alone ("would", "could", "will")
- Words that are the same but have different meanings based on context ("set" can be a verb, adjective, or noun)

Camille Mohammad Bryan

Problems with Data in Assignment 1

multiple contexts

- different meanings of words
- similar words not being counted as the same

Missing Context

- missing pages
- table of contex index etc not relevant to analysis
- ~~∞~~ words in a vaccum, no information on word frequency in relation to other words

Incorect tampered / filtered

~~Legalise~~ copyright /

- leagalise words appearing too frequently, formmatting words that have nothing to do with context of text modding results.
- Incorrect digital copies
- empty characters

Data Cleaning

→ Words with :
→ says, ~~d~~ said ?
→ Capitalization - dealt
→ non - author text ? gutenberg preamble
→ wrong author
→ wrong unit → intro, chapter?
→ translated → is it the original the same
→ why exclude stopwords
→ what about handwriting ?
→ lost word order
→ look at infrequent words
or word distribution ?

why some ~~times~~ after said not ?

Sasha will check