Text Mining Project Update 1

Planning Update

Task	**Person**
Stage 1: Preprocessing	
Preprocessing (loading the data, making any necessary adjustments etc.)	Andrea
Stage 2: Building the RSs	
Model 1: Calculation of similarity scores for “book descriptions” feature	Clarice & Anne Marijn
Model 2: Calculation of similarity scores for “book rating” feature + implementing number of “written reviews” feature as weight to the vectors	Tereza
Model 3: Combination of scores for “book description” and “book rating”	Clarice & Andrea
Stage 3: Comparison	
Our RS input: book title Our RS output: 3 times top 3 recommended book titles (1 for each model) **To-Do:** ***From our RS, we collect:*** ● The output’s 3 title ● *Manually* ⇒ The output’s books’ 3 corresponding genres as provided by Goodreads ***From Goodreads’ RS, we collect:*** ● *Manually* ⇒ The input title’s 3 corresponding genres as provided by Goodreads ● *Manually* ⇒ The 3 corresponding genres of the top 3 recommended books provided by Goodreads	Clarice, Andrea, Anne Marijn, Tereza

Week 19 (5-12 May)

- Project Update 1: Wed 8th May
- Stage 1: Pre-processing: Andrea
- Planning for models: Anne Marijn, Tereza, Clarice

Week 20 (13-19 May)

- Project Update 2: Sun 19th June
- Stage 2: Writing models: Anne Marijn, Tereza, Clarice

Week 21 (20-26 May)

- Revise code, make adjustments (using Update feedback) altogether

Week 22 (27-2)

- Deadline Final Project Sun 2nd June
- Stage 3: Comparison (altogether)
- Writing final report

Project Update

Pre-processing

- **Our Csv files:**
  - 1-100k books
  - User rating 0 to 1000
- **Merging:** Combined and matched these datasets based on book titles. This process automatically removed any books with missing ratings.
- **Renamed ID variables** for clarity as one was referring to the book ID and the other one was referring to the user rating it.
- **Handling missing values:**
  - Found 3 columns with missing values: Publisher, ISBN & Language.
  - Decided to ignore the publisher and ISBN missing values as these columns are not crucial for our genre alignment recommendation system.
  - Language directly affects user preferences and accessibility so it is an important factor in book recommendation systems. Decided to implement mode imputation for missing values in this column.
- Created a function to **tokenize and lemmatize text columns**
  - Applied function to the Name, Authors, Publishers and Rating_y (written rating) columns
- Divided data into a **training and validation set**