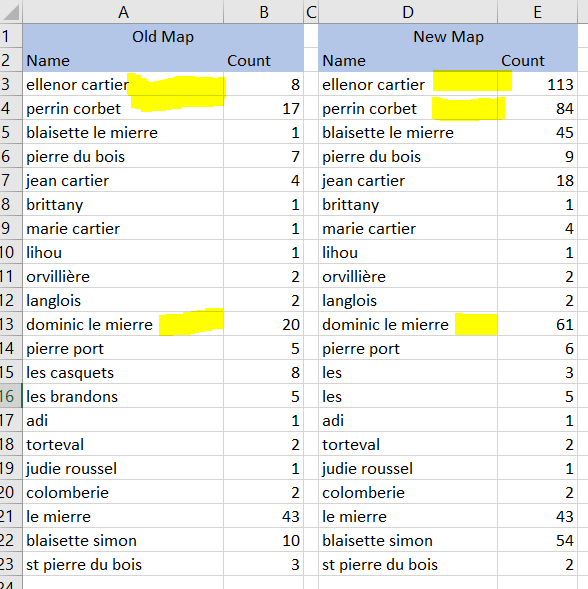
**Milestone 1:**

Books are defined by the characters that they bring to life in a reader’s mind. The presence of a protagonist to cheer for, an antagonist to pit the readers against and a medley of characters that bring the story to life are just the beginning of the beauty that the plot characters of a book can encompass.

**Extract main character :**

**Method :** In order the extract the main character, the NER(Named Entity Recognition) feature of the Stanford CoreNLP package is used. The contents of a book are sent to the CoreNLP package to be parsed. From the annotations returned, the named characters are identified as the tokens annotated with the “PERSON” tag. Each occurrence of a NE(Named Entity) is then stored in a hash map along with the number of times the entity is present in the book. Once the hash map is populated, similar keys are clubbed together using a rule based logic. For example, “Sherlock Holmes” and “Sherlock” are mapped to the same key (this has been done previously with the key “Sherlock” being removed from the map. The logic has been extended by also adding the number of occurrences of “Sherlock” to the key “Sherlock Holmes”). The figure shows the old character map as well as the new one obtained post making some experimental rule based changes. It is the map obtained for the book “Where the Deep Seas Moan” by E. Galienne Robin.



With the count of the occurrences of the main character now available in the map, the feature to be encoded at the book level is now calculated as:

(Number of occurrences of main character) / (Total occurrences of named entities in a book)

This feature would represent the weightage / impact of a main character in the book. The value returned will be normalized over the entire corpus and then added to the feature vector table. In the event that further refinement is required to the character map, the Python based dedupe library would be used.

**Results obtained on test books :**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Book Name | Main Character | No. of Occurrences | Total Character Occurrences | Ratio |
| Adventures of Sherlock Holmes | Sherlock Holmes | 97 | 1259 | 0.077045 |
| Moby Dick | Ahabs | 490 | 2741 | 0.178767 |
| Tarzan | Tarzan of the Apes | 388 | 1648 | 0.235437 |
| Emma | Emma Woodhouse | 868 | 6288 | 0.138041 |
| Where Deep Seas Moan | Ellenor Cartier | 113 | 491 | 0.230143 |

In addition to this, an intermediate character map for the entire corpus would be maintained that would be **used in Feature 3**. This map would contain all the named entities in the corpus. The reason for creating this is to be able to match against this map while performing topic modelling. The entries could also be used to refine the de-duplication step.

**Why is this method used?**

The Stanford CoreNLP pipeline has the highest accuracy when it comes to detecting names entities in a document. When paired with the rule based system for de-duplication of unique characters in a book, the results obtained were satisfactory.

**Dialog Interaction :**

**Method** **:** In order to capture dialog interactions in a book, a rule based pattern matching is used to identify words enclosed by double braces occurring near named entities. The pattern would be as follows:

\*(NER)’ ‘[word]’ ‘[word]’ ‘[word]’ ‘[word]’ ‘ “[word]\*” ‘[word]’ ‘[word]’ ‘[word]’ ‘[word]’ ‘ (NER)\*

The pattern would be adjusted to match all possible occurrences of [word] within a window of words of size 5 on either side of the double quotes. It would also be able to recognize German double quotes which may be encountered in various German books. The feature to be encoded would be the ratio of the count of the total words enclosed in double quotes to the total number of words in the book. This feature would effectively capture the author’s writing style with respect to the amount of dialogues in the book. Research works by Elson and McKeown use a similar approach to leverage the number of quotes to the total words in order to capture quoted speech as a feature in literary narrative. Hutchison and Louwerse use a co-occurrence algorithm on manually annotated data in order to retrieve the social network in a book. However, this required external annotations to be performed on the corpus in order to correctly identify the speaker and the person being spoken to in a dialogue.

**Why is this method used?**

Rule based systems are faster and less complex in terms of computation. In addition to this, other research works on dialog interaction analysis require external annotators in order to correctly identify the interactions in a dialog.

References :

1. Automatic Attribution of Quoted Speech in Literary NarrativeDavid K. ElsonandKathleen R. McKeownColumbia University
2. Gregg, Forest and Derek Eder. 2015. Dedupe. <https://github.com/dedupeio/dedupe>.
3. Manning, Christopher D., Surdeanu, Mihai, Bauer, John, Finkel, Jenny, Bethard, Steven J., and McClosky, David. 2014. [The Stanford CoreNLP Natural Language Processing Toolkit](http://nlp.stanford.edu/pubs/StanfordCoreNlp2014.pdf) In Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 55-60.[[pdf](http://nlp.stanford.edu/pubs/StanfordCoreNlp2014.pdf)][[bib](http://nlp.stanford.edu/pubs/StanfordCoreNlp2014.bib)]

**Milestone 2:**

**Documentation of results obtained during feature extraction:**

**Extract main character:**

**Results obtained:**

**Books with high main character presence:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Book Name | Main Character | No. of Occurrences | Total Character Occurrences | Ratio |
| Oliver Twist | Oliver Twist | 819 | 4207 | 0.194675 |
| Persuasion | Anne Elliott | 494 | 2816 | 0.175426 |
| Tarzan | Tarzan of the Apes | 388 | 1648 | 0.235437 |
| Emma | Emma Woodhouse | 868 | 6190 | 0.142087 |
| Where Deep Seas Moan | Ellenor Cartier | 113 | 491 | 0.230143 |

**Books with a clear main character along with multiple side characters:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Book Name | Main Character | No. of Occurrences | Total Character Occurrences | Ratio |
| Adventures of Sherlock Holmes | Sherlock Holmes | 97 | 1259 | 0.077045 |
| Memoirs of Sherlock Holmes | Watson | 119 | 757 | 0.157199 |
| Moby Dick | Ahabs | 143 | 1502 | 0.095206 |
| Emma | Emma Woodhouse | 868 | 6288 | 0.138041 |

**Books with low main character presence indicating a wide medley of characters:**

|  |  |
| --- | --- |
| Book Name | Ratio |
| Arabian Nights | 0.0802 |
| Beacon Lights of History, Volume 13 | 0.0661 |
| Real Ghost Stories | 0.0107 |
| The Valet's Tragedy, and Other Studies | 0.0242 |
| The Haunters & The Haunted | 0.026 |

**Outliers:**

Alice in Wonderland (pg11) – Main Character Presence: 0.7734

These books indicated an extremely high main character presence, due to the fact that other characters (For example – The Mad Hatter in the book ‘Alice in Wonderland’), although they were recognized by the NER tagger as named entities in a book, their occurrence counts were low compared to what one would normally expect.

**Dialog Interaction:**

For this feature, dialog interactions were identified with checking named entity presence within a window of size 7 tokens around quotes. The ratio of the number of words within quotes to the total number of words in the book is encoded as a feature.

**Results Obtained:**

**Books with high dialog interaction:**

|  |  |
| --- | --- |
| Book Name | Dialog Ratio |
| The Man Who Was Thursday: A Nightmare | 0.6158 |
| The Turn of the Screw | 0.6367 |
| Sky Island | 0.6293 |

**Books with medium dialog interaction:**

|  |  |
| --- | --- |
| Book Name | Dialog Ratio |
| The Mad King | 0.4035 |
| Emma | 0.52 |
| The Return of the Native | 0.4693 |

**Books with low dialog interaction**:

|  |  |
| --- | --- |
| Book Name | Dialog Ratio |
| Tales of Unrest | 0.2329 |
| Comic Arithmetic | 0.1117 |
| The Son of Tarzan | 0.24 |

**Results Obtained On Known Books:**

|  |  |
| --- | --- |
| Book Name | Ratio |
| Oliver Twist | 0.4473 |
| Adventures of Sherlock Holmes | 0.5052 |
| Tarzan | 0.5924 |
| Emma | 0.52 |
| Great Expectations | 0.5075 |

**Outliers:**

Certain books that contain mostly narrative text returned very low numbers for dialog interaction, which is expected. (For example – Beacon Lights of History – 0.0497).

Books written as plays also returned low numbers because the rule based capturing of dialogues using quotes did not work in this case. (For example - The Sleeping-Car: A Farce – 0.0086).

**Planned Refinements:**

1. For the character map, sometimes named entities like ‘Mozart’ and ‘Bach’ are identified as named characters. The probability is high that these occurrences are because Mozart and Bach are named in casual conversation like, ‘Mozart and Bach make great music ‘.

In order to not push these names to the character map, Named Entities with less than 3 occurrences will be removed from the map. This will also handle cases like ‘Christ’ and ‘oh Jesus!’ and will take care of them not being identified as characters in a book.