Term Frequency Recommendation Engine Metric

Matthew Scanlan

Project 1 Milestone 3

Project Overview

Text Analysis applied to a document database

Count Vectorizor application

Visualization Creation

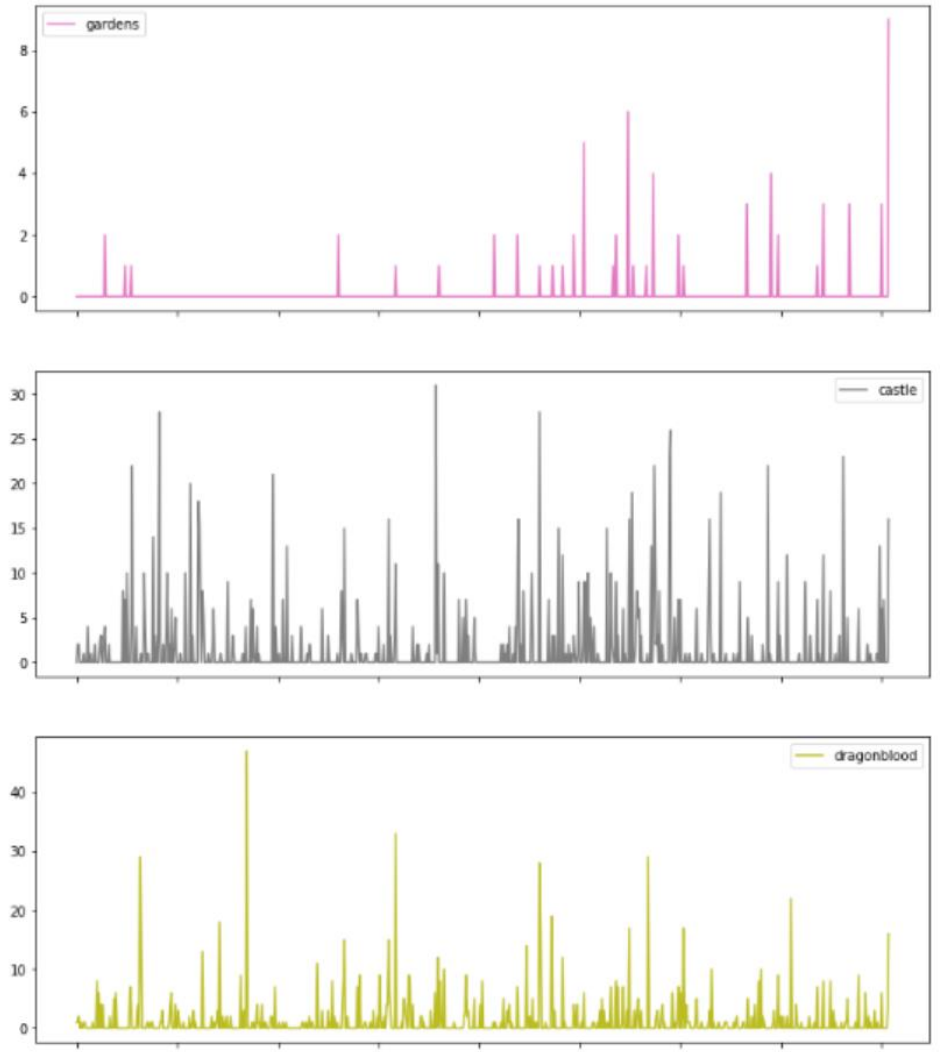
Data Description

- Sourced from website that contains web novels
- Over 800 files all formatted as .txt
- Each file contains one chapter of the novel
- Contents of each file was shuffled
- Stop words removed
- Punctuation removed

Approach

1. Read in files within document database
2. Remove Stop Words
3. Strip symbols
4. Apply TF-IDF all text within document database for one book to obtain overall top 20 words
5. Apply Count Vectorizer to each chapter
6. Specifying previously obtained top 20 words as Count Vectorizer's vocabulary
7. Utilize Count Vectorizer to generate visualization for end user review

Sample word visualization



Information contained

- Where the terms occur within the document database
- High level overview of the subject of the document database or book
- Metrics relevant or useful for a recommendation engine

Application

- Recommendation engine metric
- Visualization provided to user along with recommendation from search engine

1

Generalizing for application to all document databases searched by recommendation engine

2

Running model when document database is added to recommendation engine's scope to optimize end user experience

Recommendations