Sentiment Analysis Data Pre-processing

Anindya Dutta and Aarushi Goel

September 18 2017

1 Introduction

Project Gutenberg offers over 54,000 e-books. We are choosing a subset of a few hundred books for the purpose of this assignment. For simplicity, we are not considering anthologies as they may have conflicting results depending on different stories.

1.1 Bookshelves

For the first phase, we will use books from five bookshelves (genres). Though genre analysis is not the primary goal of this task, we will store this data for future reference. The bookshelves are **children, classic, crime, history,** and **psychology**.
We have chosen the above categories because we think that these are wide genres, and that it will be a lesser probability of misclassification because they are quite disjoint.

2 Data Pre-processing

2.1 Downloading the e-books

Every e-book is available in HTML, ePub and Plain Text UTF-8 formats. We will be using the plain text format.

2.2 Feature set

The feature set for each book would have the **ID, text** and **sentiment** associated with the book.

1	Mariam was eigth...	grief
2	Mr. and Mrs. Dursley...	happy
3	...	...

Additionally, we will save three meta-features that may be required for use in the future. They are the **author, genre,** and **year published**.

2.3 JSON structure

The JSON for one e-book can then look like Listing 1.

3 Sentiment Analysis

Sentiment analysis in the basic form can be done by classifying words as positive or negative. We have currently downloaded lists of positive and negative words. We are currently reviewing NLTK Sentiment Analysis package to see what algorithms already exist and how we can utilize them.

```
{
    id: "ISB..."
    text : "The morning had dawned clear..",
    sentiment : "sad"
    meta : {
                title: "A Game of Thrones"
                author: "George R.R Martin"
                genre: "Fantasy"
                year: 1997
            }
}
```

Listing 1: JSON for book