BOĞAZİÇİ UNIVERSITY

Cmpe 493

Assignment 3 - Report

Arzucan ÖZGÜR

Baran Deniz KORKMAZ

FALL 2020

1 Introduction

In this assignment, we are asked to implement a book recommendation system using Goodreads database. The recommendation system is expected to apply content based filtering.

2 Program Interface

The program has been written in **Python 3**. The program has been tested in **Python 3.6.9**. The following standard-library modules are used:

1. os
2. sys
3. pickle
4. string
5. re
6. math
7. shutil
8. urllib

3 Program Execution

A query can be given into the program in two steps. In order to run the first step, the program assumes that **stopwords.pickle** is provided in the working directory.

1. Fetch and save the contents of the books, build, and save the model:

   ```
   >>> python3 main.py [path_to_file]
   ```

   NOTE: The program saves the content of the books in the file **metadataAss3.pickle**. **NOTE:** The program saves the models built in the folder **Assignment3Model**.

2. Now that we have already constructed the required structures, enter the following command to process a query:

   ```
   >>> python3 main.py [url]
   ```

Example:

```
>>> python3 main.py https://www.goodreads.com/book/show/18050143-zero-to-one
```

4 Program Structure

This section describes the main stages of the program.

4.1 Data Preprocessing

The following steps have been performed for data preprocessing:

4.1.1 Scraping the Goodreads

In this step, the content of the books embedded in the HTMLs of the website have been scraped in the desired format using **urllib** for fetching and **RegEx** for data extraction.

4.1.2 Tokenization

In this step, the plain text content of title and body has been tokenized in a way that each word has been divided by whitespaces.

4.1.3 Normalization

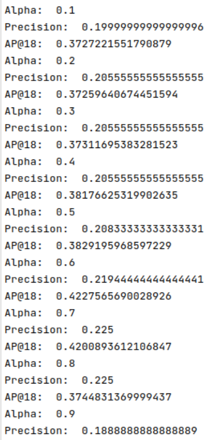
In this step, the following operations have been applied to tokens obtained at the previous stage:

1. Case-Folding
2. Punctuation Removal
3. Stopword Removal (Using the stopwords list provided by **nltk**)

We must note that after the punctuations are replaced by a single whitespace, the tokens have been retokenized in a way that there exists no whitespaces in the final tokens.

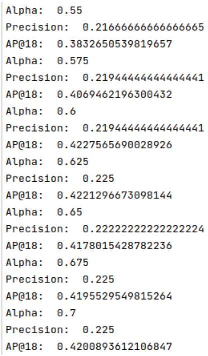
4.2 Model Details

The fields of books that will be utilized for building recommendation system are descriptions and genres. In the assignment, I have used TF-IDF weighting scheme for implementing a vector-space model that will be used to analyze the similarity between pair of books. After the data preprocessing, the terms that we obtain form the vocabulary. We represent the books and the query as vectors based on the TF-IDF weights that are computed using the aforementioned vocabulary. In the constuction of model, I have not implemented any mechanism to select the informative terms as the evaluation results obtained during testing stage were promising. However, I have analyzed the effect of alpha that will determine the weights of similarities of descriptions and genres for book and query pairs. For this purpose, I implemented a cross-validation like mechanism that will output the precision and AP@18 for different predetermined alpha parameters. I tested the alpha values starting from 0.1 to 0.9. Then I picked the most promising interval of [0.55,0.7]. My observations for a subset of 20 and 100 books given as query suggested that the most optimal value for alpha is 0.625. Below you can see the illustrative output examples for the optimization stage for parameter alpha:



```
Alpha:  0.1
Precision:  0.19999999999999996
AP@18:  0.3727221551790879
Alpha:  0.2
Precision:  0.20555555555555555
AP@18:  0.37259640674451594
Alpha:  0.3
Precision:  0.20555555555555555
AP@18:  0.37311695383281523
Alpha:  0.4
Precision:  0.20555555555555555
AP@18:  0.38176625319902635
Alpha:  0.5
Precision:  0.20833333333333331
AP@18:  0.3829195968597229
Alpha:  0.6
Precision:  0.21944444444444441
AP@18:  0.4227565690028926
Alpha:  0.7
Precision:  0.225
AP@18:  0.4200893612106847
Alpha:  0.8
Precision:  0.225
AP@18:  0.3744831369999437
Alpha:  0.9
Precision:  0.18888888888888889
```

Figure 1: Cross-Validation 1



```
Alpha:  0.55
Precision:  0.21666666666666665
AP@18:  0.3832650539819657
Alpha:  0.575
Precision:  0.21944444444444441
AP@18:  0.4069462196300432
Alpha:  0.6
Precision:  0.21944444444444441
AP@18:  0.4227565690028926
Alpha:  0.625
Precision:  0.225
AP@18:  0.4221296673098144
Alpha:  0.65
Precision:  0.22222222222222224
AP@18:  0.4178015428782236
Alpha:  0.675
Precision:  0.225
AP@18:  0.4195529549815264
Alpha:  0.7
Precision:  0.225
AP@18:  0.4200893612106847
```

Figure 2: Cross-Validation 2

4.3 Query Processing

In the final step, given that the required preprocessing and construction of data structures have been satisfied, we are ready to process a given query. Below, you can see an example output for the url `https://www.goodreads.com/book/show/18050143-zero-to-one`.

```
Processing URL id:  1
URL: https://www.goodreads.com/book/show/18050143-zero-to
    -one Status Code: 200

The content of the book for query: https://www.goodreads.
    com/book/show/18050143-zero-to-one
Book URL: https://www.goodreads.com/book/show/18050143-
    zero-to-one
Title: Zero to One: Notes on Startups, or How to Build
    the Future
Authors: Peter Thiel, Blake  Masters
Description: If you want to build a better future, you
    must believe in secrets.The great secret of our time
    is that there are still uncharted frontiers to explore
     and new inventions to create. In Zero to One,
    legendary entrepreneur and investor Peter Thiel shows
    how we can find singular ways to create those new
    things. Thiel begins with the contrarian premise that
    we live in an age of technological stagnation, even if
      were  too distracted by shiny mobile devices to
    notice. Information technology has improved rapidly,
    but there is no reason why progress should be limited
    to computers or Silicon Valley. Progress can be
    achieved in any industry or area of business. It comes
     from the most important skill that every leader must
    master: learning to think for yourself.Doing what
    someone else already knows how to do takes the world
    from 1 to n, adding more of something familiar. But
    when you do something new, you go from 0 to 1. The
    next Bill Gates will not build an operating system.
    The next Larry Page or Sergey Brin  wont  make a
    search engine. Tomorrows champions will not win by
    competing ruthlessly in  todays  marketplace. They
    will escape competition altogether, because their
    businesses will be unique. Zero to One presents at
    once an optimistic view of the future of progress in
    America and a new way of thinking about innovation: it
     starts by learning to ask the questions that lead you
     to find value in unexpected places.
```

```
Recommendations: https://www.goodreads.com/book/show
    /10127019-the-lean-startup, https://www.goodreads.com/
    book/show/4865.How_to_Win_Friends_and_Influence_People
    , https://www.goodreads.com/book/show/368593.
    The_4_Hour_Workweek, https://www.goodreads.com/book/
    show/18176747-the-hard-thing-about-hard-things, https
    ://www.goodreads.com/book/show/2612.The_Tipping_Point,
     https://www.goodreads.com/book/show/25541028-elon-
    musk, https://www.goodreads.com/book/show/7723797-
    business-model-generation, https://www.goodreads.com/
    book/show/22668729-hooked, https://www.goodreads.com/
    book/show/11468377-thinking-fast-and-slow, https://www
    .goodreads.com/book/show/36072.
    The_7_Habits_of_Highly_Effective_People, https://www.
    goodreads.com/book/show/13078769-running-lean, https
    ://www.goodreads.com/book/show/1134122.
    The_Power_of_Positive_Thinking, https://www.goodreads.
    com/book/show/98233.Founders_at_Work, https://www.
    goodreads.com/book/show/763362.The_One_Minute_Manager,
     https://www.goodreads.com/book/show/69571.
    Rich_Dad_Poor_Dad, https://www.goodreads.com/book/show
    /43848929-talking-to-strangers, https://www.goodreads.
    com/book/show/11084145-steve-jobs, https://www.
    goodreads.com/book/show/30186948-think-and-grow-rich
Genres: business, nonfiction, entrepreneurship, economics
    , selfhelp, technology, management, leadership,
    buisness, finance

Predicted Recommendations:
1) Blue Ocean Strategy: How to Create Uncontested Market
    Space and Make the Competition Irrelevant - W. Chan
    Kim, Ren e Mauborgne
2) Built to Last: Successful Habits of Visionary
    Companies - James C. Collins, Jerry I. Porras
3) The E-Myth Revisited: Why Most Small Businesses Don't
    Work and What to Do About It - Michael E. Gerber
4) The Lean Startup: How Today's Entrepreneurs Use
    Continuous Innovation to Create Radically Successful
    Businesses - Eric Ries
5) The Hard Thing About Hard Things: Building a Business
    When There Are No Easy Answers - Ben Horowitz
6) Good to Great: Why Some Companies Make the Leap... and
     Others Don't - James C. Collins
7) Rework - Jason Fried, David Heinemeier Hansson
8) The Four Steps to the Epiphany: Successful Strategies
    for Startups That Win - Steve Blank
```

```
9) Start with Why: How Great Leaders Inspire Everyone to
    Take Action - Simon Sinek
10) The Leader Who Had No Title: A Modern Fable on Real
    Success in Business and in Life - Robin S. Sharma
11) Business Model Generation - Alexander Osterwalder,
    Yves Pigneur
12) The One Minute Manager - Kenneth H. Blanchard,
    Spencer Johnson
13) Running Lean: Iterate from Plan A to a Plan That
    Works - Ash Maurya
14) The 4-Hour Workweek - Timothy Ferriss
15) Elon Musk: Tesla, SpaceX, and the Quest for a
    Fantastic Future - Ashlee Vance
16) Founders at Work: Stories of Startups' Early Days -
    Jessica Livingston
17) Onward: How Starbucks Fought for Its Life without
    Losing Its Soul - Howard Schultz, Joanne Gordon
18) It Happened In India: The Story of Pantaloons, Big
    Bazaar, Central and the Great Indian Consumer -
    Kishore Biyani, Dipayan Baishya

EVALUATION
Alpha: 0.625
Precision: 0.4444444444444444
AP@18: 0.37948926073926076
```

5 References

- For Stopwords: https://www.nltk.org/index.html