Intermediate Project Report
Maria Antoniak and Antariksh Bothale
LING 575 Sentiment, Subjectivity, and Stance
May 13, 2014

PROJECT GOALS

Our goal is to classify books into genres based on aspects extracted from user reviews. We hypothesize that readers of certain genres have concerns particular to that genre, e.g. a reader of romance novels might be more concerned with the protagonist's love interest while a reader of non-fiction might be more concerned with the author.

More specifically, we plan to use an LDA package to extract aspects from book reviews and then use those aspects as features for a classification package that will assign genres to the books.

MAJOR CHALLENGES

The first major challenge we face is gathering data. We were not able to identify a single resource or corpus for our needs, so we need to blend two resources together. See "Project Resources" below.

Another challenge is finding software to perform the aspect extraction. We have not been able to locate an open source LDA package that allows seeds, so we have scaled back our approach to just use a standard LDA (either MALLET or Gensim). In retrospect, this might be a better approach for our purposes, since we are not trying to summarize the aspects for human consumption but rather feed those aspects into a classification algorithm that will assign genres to the books.

Perhaps most importantly, we are unsure how well our approach to genre classification will work. Perhaps using aspects as features will work for some genres and not others, or perhaps aspects will only be helpful in conjunction with other features. This will require some experimentation and feature engineering as we near the end of our project.

CURRENT APPROACHES

We have been unable to identify any previous work that attempts our specific task (genre classification based on user reviews). However, there is ample work in each of our subtasks and especially in aspect extraction.

In particular, we are interested in Arjun Mukherjee and Ben Liu's paper *Aspect Extraction through Semi-Supervised Modeling*. This paper gave us our first inspiration in using aspect extraction for our task. We were initially interested in using the same methodology (a seeded version of LDA) for our aspect extraction, but unfortunately we have not found a free implementation and do not have time to rewrite the LDA modeling ourselves. However, we have taken some perspective from this paper and will use an LDA, albeit unseeded.

LIMITATIONS AND OUR WORK

One major limitation that we foresee is that there might not be significant variation of key aspects and talking points across the genres, or that this method might not be successful in picking out genres beyond very obvious ones with stark differences (say romance v/s non-fiction). Additionally, since we are following an unsupervised approach, the algorithm will just produce cluster, and we foresee some difficulty in assigning those clusters to specific genres.

PROJECT RESOURCES

One of the first challenges we faced in this project was identifying a resource that combined all the data we required. We were unsuccessful, and so we have decided to manually combine data from two different resources. We are gathering user reviews from the SNAP dataset of Amazon reviews. We filtered these reviews for ones that are about books and that contain ISBNs, and we then use the ISBNs to scrape genres from GoodReads.com. So far, this approach has been successful, though the scraping is quite slow.

For the aspect extraction, we are planning on using LDA modelling. We could not find a resource that exactly matched what we were hoping for, so we have fallen back on using either MALLET or Gensim for this task.

We have not yet decided what algorithm/model to use for the final clustering/classification of books by genre. We will probably use a package like MALLET for this task.

We have read follow-up papers from both Mukherjee and Liu, as well as investigated incorporating parts of the Stanford Core NLP resources.

PROJECT PLANS & CURRENT STATE

We have gathered the user reviews and filtered them by product (only books) and ISBN (some book reviews do not contain ISBNs). We then scraped ranked lists of genres from GoodReads.com for each unique ISBN. We have gathered genre information for about 15,000 books, and we are currently optimizing this process, since it has taken a couple days to reach this point and we would like to gather data for about 300,000 books.

Our next step is transform the data into vectors for MALLET or Gensim. We will then feed this through an LDA package and examine the resulting aspects. Finally, we will incorporate these aspects into a clustering or classification algorithm to assign genres to each book.

We have not yet established a baseline, since this is a three-step process (data gathering, LDA aspect extraction, genre classification) and we have only just finished the first step. We hope that the second and third steps will move fairly quickly and that we will then have time to play with the features and aspects to improve our results.