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.

CURRENT APPROACHES

LIMITATIONS & OUR WORK

PROJECT RESOURCES

PROJECT PLANS & CURRENT STATE

* Project goals: What is the main task you are tackling or problem you are investigating?
* Major challenges related to your project: What makes it hard?
* Current approaches: What theories have been proposed to explain these phenomena? What techniques have been applied to solve these challenges? You should have identified at least one paper to start your investigation of your topic and refine your ideas.
* Limitations and your work: What problems do you see in these proposals? How will you test these theories or algorithms?
* **Project resources:** What corpora or datasets will you use to assess these approaches? What software or toolkits would be helpful? What other papers have you chosen to study and evaluate?
* Project plans and current state: What steps have you taken so far? Have you established a baseline or preliminary implementations?