Hannah Gray
Ling 294-02
Stephanie Farmer
5/7/16

Final Project: German Sentiment Analysis

Introduction

As a student studying German here at Macalester, I was interested to try and see if I could try my hand at building a sentiment analyzer for German words and then use that analyzer to predict the sentiment for different sentences in German. I wanted to see what German words were considered most strongly negative or positive, and then see if they could provide an accurate depiction of what the rest of the sentiment of a sentence was. As my project progressed, this evolved into more specifically sentences about film, however the future end goal would be something more universal.

Background

There has been lots of literature about sentiment analysis done by the IGGSA (Interest Group for German Sentiment Analysis. They have created a corpus of 270 sentences manually annotated for objectivity and subjectivity, word and phrase polarity and expressions of private states. A lot of their research has to do with data mining of sentiment. However, in 2011 they published a paper about positivity and negativity in hotel reviews, which provided a touch point for my project since I ended up doing a similar--albeit much more simple--project for film reviews.

Design

My project has two major components: The data collection aspect and the data analysis aspect. In the data collection part, I scraped the text from the german film review website filmstarts.de. I took the top 20 movies in theaters right now, and took the user reviews from the first page of those ratings--eventually I would have liked to do every page of ratings for every movie, but I found it was quite a lot of work to parse through. The movies I used were: Ein Mann namens Ove, Jungle Book, Avengers Civil War, How to be Single, Zoomania, the Boss, Rico, Oskar und der Diebstahlstein, Gods of Egypt, Ein Hologramm fur den König, Birnenkuchen mit Lavendel, Bad Neighbors 2, The Huntsman and the Ice Queen, Bib and Tina: Mädchen gegen Jungs, Triple 9, Batman vs. Superman, Die Kommune, Rachet and Clank, Hardcore, The Lady in the Van, and Kung Fu Panda 3.

After getting the movies and finding each review and star rating for all of those ratings, cleaned up the text, took out stopwords, and assigned each word a positivity or negativity rating between -3 and 2 (-3 = 0 stars, 2 = five stars). I don't know if it was the best idea to take out the stop words, but I didn’t want all the itty bitty words like *der*, *die,* or *das* (all forms of the article ‘the’ in English) to skew the positivity/negativity of a sentence just by being there.

The second major component of my script was the data analysis side. Here, is where I applied the dictionaries of words that I had collected earlier, and used those words to decided on the overall positive or negative sentiment of another input (originally I was hoping to use tweets from twitter, but I was having trouble webscraping German tweets, so I used my own inputs instead). The input sentence was turned into unicode (to make sure it was compatible with the words from the review) and then broken into its individual words. I then compared each word to the words in the positive dictionary and the negative dictionary and assigned those sentiment values to the words in my sentence. I then averaged the positive and negative values of the words in the sentence and gave it an overall sentiment rating depending on which value was higher.

Concerns

My corpus of data is very small, because I only took the first page of reviews from the top 20 movies in theaters right now. This could be addressed by going through and adding code to web scrape the later pages of movies with more than one page of reviews. I could also add more movies to my dataset to increase the corpus size. Most of that is limited by time constraints. Movies on filmstarts.de are labeled by codes (example, 223207 is the code for Zootopia) and thus I have to find the codes that they use for each movie, instead of just typing them in myself.

Unfortunately, the smallness of my corpus means that most words only have one or two instances of repeat in my data set and thus, their positivity or negativity comes from that one review. More data would help me get a more precise value for different words.

Also, my positive and negative words include all of the different verb forms for individual verbs, as well as different adjectives with different endings etc. which makes it harder to get an exact representation of the total positivity/negativity of a base word. I would have liked to find a way to just have the single word form (although it is possible there are some interesting conclusions to be found if perhaps *groß* [large, great] and *größer* [larger, greater] had different positivity counts since one might appear in more positive reviews than the other).