Georgios Akritidis
Department of Media Technology
Linnaeus University
ga222ey@student.lnu.se

Introduction

This document is the description of my final project proposal for the course Programming for Digital Humanities. The final project is about sentiment analysis which refers to the use of natural language processing and text analysis, in order to identify what is the opinion of the people that write the text that is being analyzed; which could be a comment in a web page, a tweet or an article. The idea is that through the analysis of the text, we can make assumptions about how positive or negative is the opinion of the people that wrote this text, concerning a specific topic. In my project, my task will be to examine the opinion of the people regarding the movie Star Wars: The last Jedi. Specifically, I will collect a substantial amount of tweets relative to this matter (100-200 tweets), analyze their polarity and present the findings. My goal is to find out whether the promotion of this film and the good criticism that has received, corresponds to reality and if it does not how much the real popularity differs from what is presented by the film critics..

Methodology

The main tool that I will use to implement my project is textblob which is a Python library for processing textual data. It is easy to install and it is equipped with commands that make it easy to find the polarity (level of popularity) of a given text. Before, I start with writing the code I will have to sign up in twitter developers and then create an application, in order to obtain an api key, api secret, a token or a token secret. The results of the code will be stored in a csv file and similarly as I did for the assignment 3 I will create graphs and charts to visualize the findings. Furthermore, I will have to install the api tweepy in my system which will allow me to collect the tweets.

Also, I will have to import json get the data and additionally I will import the library mathplotlib so that I can create plot directly after executing the code. The reason for the embedding of plot is that I want to observe the visualization of the data instantly so that I can make immediately changes to my code when necessary. Basically, in every figure there will be three values (positive, neutral, negative) and each one of them will be presented with a different color. To be more accurate, positive opinions will take the green color, neutral the blue color and negative the red. After, I decide if the results are what I wanted, then I will

proceed on creating extra graphs with the use of commercial web sites.

The reason I prefer textblob for the implementation of my project is that this library has ready commands that produce the polarity of a text. So, it gives me the possibility to produce the results I want quickly and effectively. Mathplot is a very useful library that gives the possibility to generate figures with labels embedded on them, such as the x,y axis, the title e.t.c. Also, it allows to select different colors to represent the values of the dataset.

Major blocks of code

The different programming topics that will concern my final project will be related to text processing, sentiment analysis, and data visualization. Specifically, my aim is to present the most used adjectives that indicate opinion and feelings about the movie and present them sorted from highest to lowest. Then, I will present the values of polarity for every comment, whether the opinions are positive or negative. Finally, I will create multiple graphs so that the reader has better view of the findings.

Expected outcome

After executing the code, in the command line, there will be displayed the comments, the name of the user, the polarity and the subjectivity. At the same time, a figure will be generated that will demonstrate the number of tweets and the range of polarity. In addition, a csv file with four columns will be also generated. The first column will be a text column with the actual tweet, the second another text column with a value either positive, neutral or negative. The other two will present the values of polarity and subjectivity. With this file, I will create two other graphs, a scatter plot and a dentrograph.