Fall 2013

NLP/ML/WEB by Pr. Andrew Rosenberg

Linguistic Oriented Requirement

Introduction: Response to document representation

In this homework we chose to work on classifying webpages from the web site Metacritic. Metacritic is a site that focuses on summarizing entertainment reviews for all types of media, specifically games, movies, TV, and music. The web document we focus on is Metacritic's media review page. Each category of media has a breakdown of every item ever reviewed and each item is given its own review page. For example, for the category movies every movie listed will have its own personal review page that contains basic information, a metascore, a list of user reviews and critic reviews, and more. An example of one media review page can be viewed at www.metacritic.com/movie/don-jon. Using this information the 5 categories we chose for classification are listed below:

(1) Media name (TV, game, music, movie)

(2) Grade (Good, average, bad)

(3) Old/New (released 2000-2003 or released 2010-2013)

(4) Score (metascore between 1-100)

(5) User ratings (total number of reviews given by users)

We chose 3 distinct document representations. In the following section, section 1, we give a breakdown of the methods we used. In sections 2 and 3, we give a break down of our responses to each word representation. We combined the discussion of bigram and trigram representations to one section, section 3. Our last section, section 4, discusses possible improvements and what we feel would be an ideal word representation for our task.

Section 1: Methods

We began by manually balancing our training set to ensure that our training data included a balanced number of examples. Then we organically scraped the metacritic website performing a random walk using our web crawler. Labeling was then done automatically. The features we extracted were the strings of text pulled from each webpage. We obtained our representations from these features. The breakdown of our three document representations is explained below.

Section 2: Bag-of-words

Our first representation was a bag-of-words representation. Using the tool Weka we were able to train our classifier using the sequential minimal optimization algorithm and then run tasks on our test data. We hypothesized that this representation would be useful in identifying the media name and grade. These categories would tend to have keywords that would help to easily distinguish between labels. For example, it would be more common for the word *show* to appear with TV or the word *album* to appear with music. But since this representation lacked ordering information we felt it would not be as useful in the other categories. We were correct in our predictions using this representation we were able to classify correcting 100% of the media names instances, 81.5% of the grade instances, and 77.9% of the old/new instances. Our score category achieved a correlation coefficient of about 0.75 and the number of ratings achieved about a .05. We knew to improve in many of these categories we needed to use a document representation that could capture more information. Therefore, we turned to N-gram representations.

Section 3: N-gram Representations (bigram and trigram)

We felt the text was the richest way of representing our data. To improve upon our results we needed to choose a representation that would provide more data. Due to the type of document analyzed the written text could be extremely useful in determining the nature of the review. Is this a positive or negative review? Is this review page about a movie or a TV show? Patterns in text could be useful in garnering answers to these questions. Giving thought to the type of document we grappled with the question, what text is relevant to classifying this document? Common practice suggests the removal of stop words. However, considering the common prose of a review we felt stop words, in our situation, served an important role. Specifically, if we only focus on content words we may misunderstand a message that could be elucidated through the addition of conjunctions or prepositions. It would not be enough to just consider frequent words or content words for a review. For a review to be truly captured the context in which these words occur is essential. To ensure a good balance between capturing the local ordering information of the text and sparse observations we chose to use both bigram and trigram representations. We were interested in seeing how the results would change as we transitioned from a bi-gram to a trigram representation.

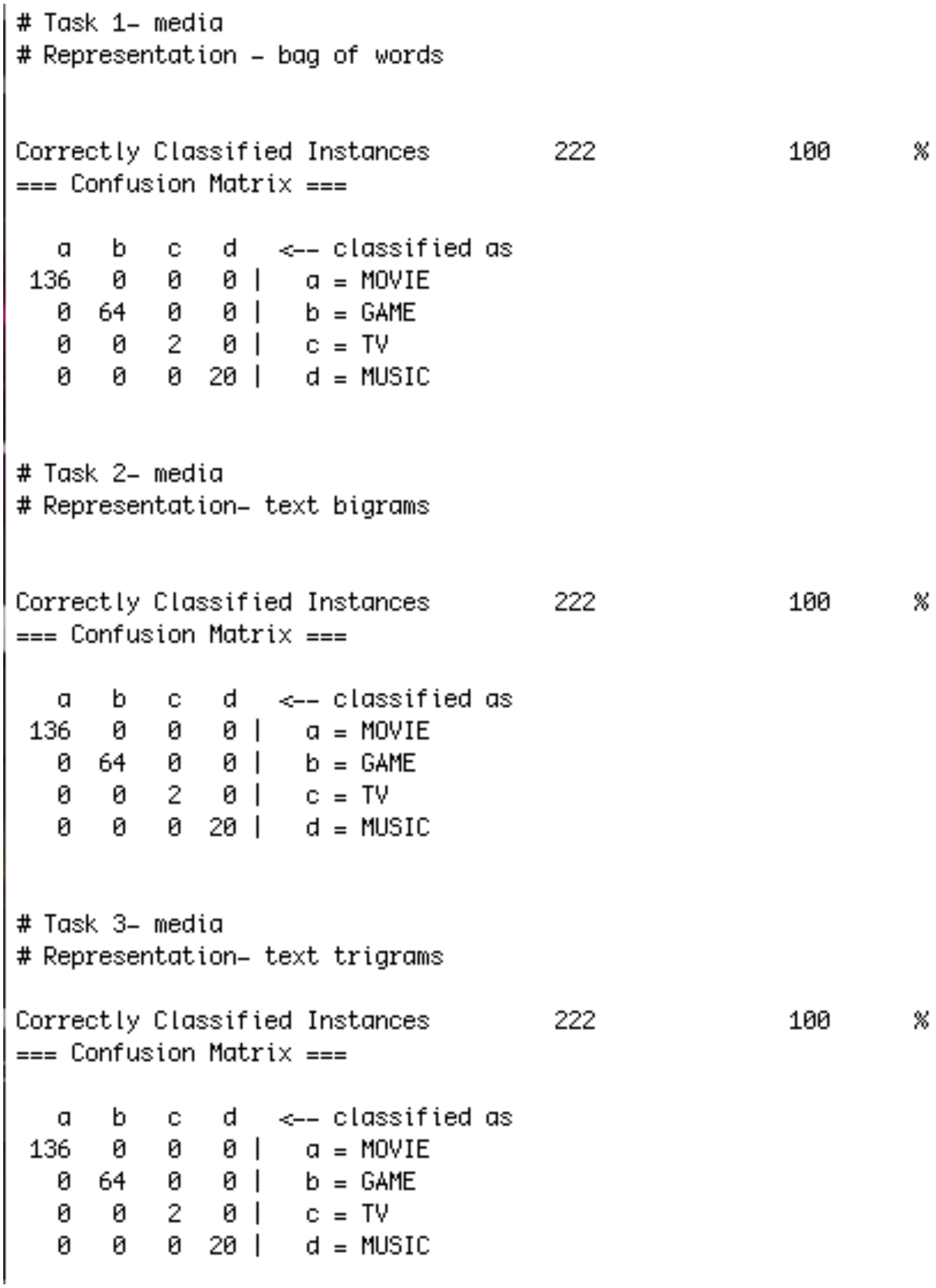
In 4 out of the 5 categories we witnessed some form of improvement using an N-gram representation. The only category that did not improve was the media name category, which was able to classify 100% of the instances using all three types of word representation. The grade category saw the most improvement and increase of about 12.6%. This improvement demonstrates the importance of ordering information in ascertaining the nature of a text document. A more detailed breakdown of results can be viewed in the appendix section.

Section 4: Ideal word representation

There does not exist one ideal word representation for all 5 of our categories. Each category is unique in nature and would thus require its own appropriate representation. In this section we choose to focus on solely one category, the grade category. We base this choice off complexity. This category appears to us to be the most interesting because it is a question of sentiment analysis. Current literature suggests that there exist many proficient approaches to sentiment analysis. Our ideal representation for the specific task of categorizing a review page, as good/average/bad would make 2 large improvements using ideas gained from current work in Natural Language Processing.

The first improvement would be the inclusion of a sentiment lexicon. A sentiment lexicon could be described as a list of words that are marked as either having a positive, neutral, or negative sentiment. A sentiment lexicon could be employed to identify words that suggest a review is of a certain kind. It could be especially useful in identifying those grey area reviews that are hard to place. Identifying the hard to place reviews would help to improve on our overall performance. The second improvement we would make would be to include a dependency parsed representation. By using a dependency parsed representation we can improve upon our representation by including more information about the text. Not only would we be able to capture the ordering of the words but also the syntactic relations between the words. By capturing the relations between words we can gain a better sense of meaning. A syntactic representation would include information about negation, conjunctions, etc. These relations are key to deciphering misleading or complex sentences. A syntactic representation would help to give the appropriate context and thus a better representation of the text. By employing these two improvements we would hope to improve upon our current system.

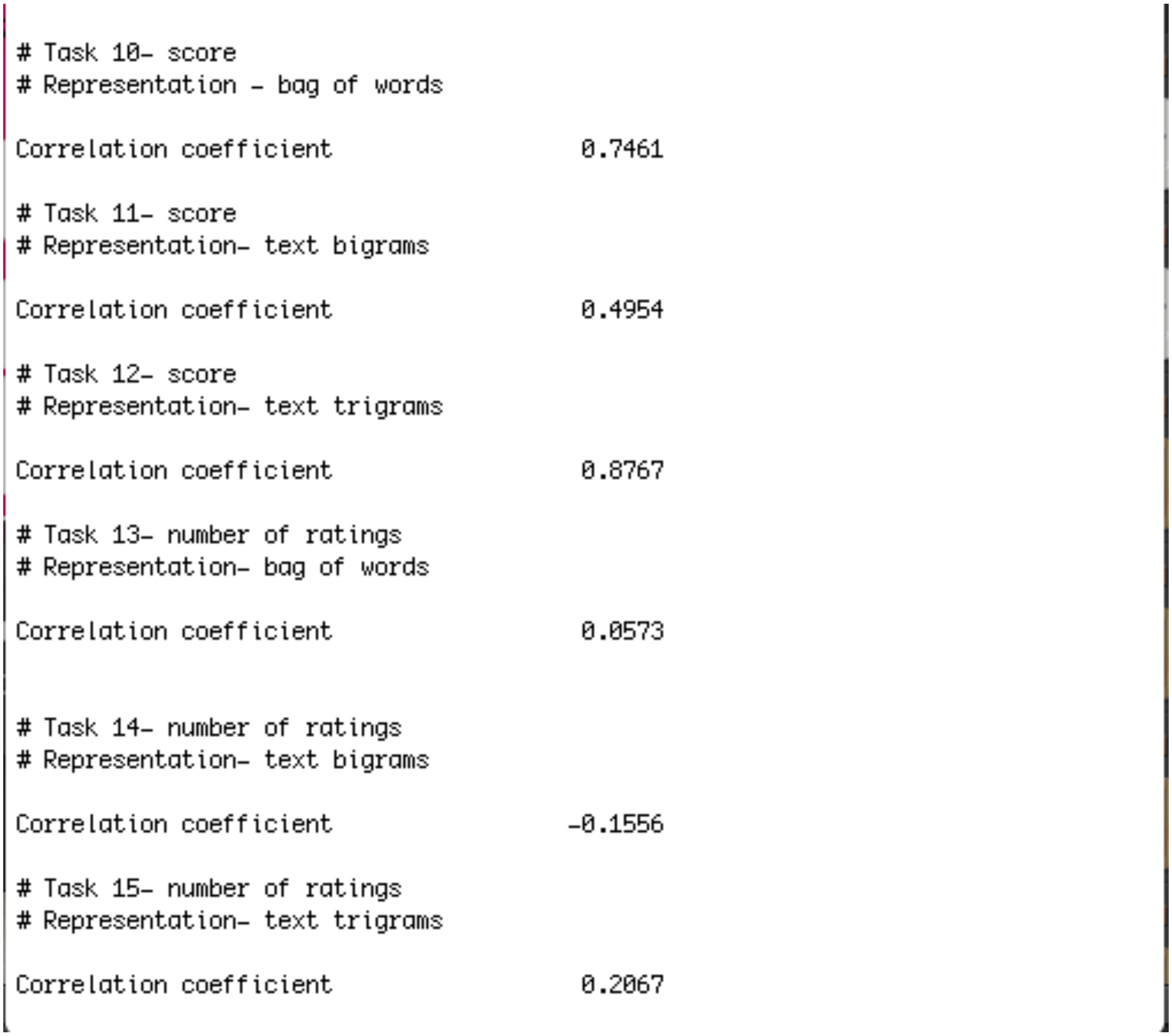
Appendix (Results)



```
# Task 1- media
# Representation - bag of words


Correctly Classified Instances         222               100      %
=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 136   0   0   0 |   a = MOVIE
   0  64   0   0 |   b = GAME
   0   0   2   0 |   c = TV
   0   0   0  20 |   d = MUSIC


# Task 2- media
# Representation- text bigrams


Correctly Classified Instances         222               100      %
=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 136   0   0   0 |   a = MOVIE
   0  64   0   0 |   b = GAME
   0   0   2   0 |   c = TV
   0   0   0  20 |   d = MUSIC


# Task 3- media
# Representation- text trigrams

Correctly Classified Instances         222               100      %
=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 136   0   0   0 |   a = MOVIE
   0  64   0   0 |   b = GAME
   0   0   2   0 |   c = TV
   0   0   0  20 |   d = MUSIC
```

```
# Task 4- grade
# Representation - bag of words


Correctly Classified Instances         181               81.5315 %
=== Confusion Matrix ===

   a   b   c   <-- classified as
 109  21   3 |   a = GOOD
   9  52   1 |   b = AVERAGE
   3   4  20 |   c = BAD

# Task 5- grade
# Representation- text bigrams


Correctly Classified Instances         209               94.1441 %
=== Confusion Matrix ===

   a   b   c   <-- classified as
 131   2   0 |   a = GOOD
   4  58   0 |   b = AVERAGE
   1   6  20 |   c = BAD


# Task 6- grade
# Representation- text trigrams

Correctly Classified Instances         209               94.1441 %
=== Confusion Matrix ===

   a   b   c   <-- classified as
 130   3   0 |   a = GOOD
   1  61   0 |   b = AVERAGE
   0   9  18 |   c = BAD
```

```
# Task 7- old/new
# Representation - bag of words


Correctly Classified Instances         173               77.9279 %
=== Confusion Matrix ===

   a   b   <-- classified as
   1   3 |   a = OLD
  46 172 |   b = NEW


# Task 8- old/new
# Representation- text bigrams

Correctly Classified Instances         171               77.027  %
=== Confusion Matrix ===

   a   b   <-- classified as
   2   2 |   a = OLD
  49 169 |   b = NEW

# Task 9- old/new
# Representation- text trigrams

Correctly Classified Instances         182               81.982  %
=== Confusion Matrix ===

   a   b   <-- classified as
   3   1 |   a = OLD
  39 179 |   b = NEW
```



```
# Task 10- score
# Representation - bag of words

Correlation coefficient                  0.7461

# Task 11- score
# Representation- text bigrams

Correlation coefficient                  0.4954

# Task 12- score
# Representation- text trigrams

Correlation coefficient                  0.8767

# Task 13- number of ratings
# Representation- bag of words

Correlation coefficient                  0.0573


# Task 14- number of ratings
# Representation- text bigrams

Correlation coefficient                 -0.1556

# Task 15- number of ratings
# Representation- text trigrams

Correlation coefficient                  0.2067
```