Clustering of Articles by Theme and Sentiment

Natural Language Processing Final Report

Courtney Fiala and Anders Maraviglia

December 9, 2016

Abstract

To objectively determine the semantic relationship between different selections of text has proven itself to be a difficult problem. Specifically, a computational method for identifying and quantifying the bias between news articles written on the same theme has yet to be implemented. Methods for sentiment and thematic extraction of textual articles have been previously explored. However, these algorithms have not been applied together in a comprehensive application to form any relationship between articles. We propose a novel solution that leverages natural language processing methods including theme extraction and sentiment analysis to graph a set of articles into clusters based on their themes and to convey the sentimental differences between articles.

1 Introduction

There are many existing methods that classify articles based on broad themes and still more methods that determine the sentiment of a particular sentence. Most theme extraction algorithms classify an article into one of a predetermined set of classes, such as politics, sports, or weather. We wish to identify more in-depth themes such as the most recent presidential debate or the championship basketball game. Further there exist a few websites that present a set of articles with opposing viewpoints on a particular subject, but these sites use hand-selected data instead of employing natural language processing methods.

A group of RCOS students is working on a project (Opinionated) to create a web application to display two news articles with differing viewpoints on a randomly generated theme. This group would like to use a computer algorithm to select which articles to display. There is currently no method to determine the themes of a set of articles then compare a pair of articles based on their sentiments towards one theme.

Our goal this semester was to create an in-depth theme extraction and sentiment analysis approach for the article clustering. This approach was to be exported into a module that can be used by the Opinionated team in their application to quantify relationships between article viewpoints.

Feature	Description
Entities	List of entities extracted using named-entity recognition
Keywords	List of keywords extracted
Taxonomy	5-level categorical theme extraction with a corresponding weight
Sentiment	Positive, Neutral, or Negative with a corresponding weight
Emotions	Weights for levels of Anger, Joy, Fear, Sadness, and Disgust
Relevance	Relevance of an entity, keyword, or taxonomy in a sentence

Table 1: Description of all features used in article comparison. Note that Sentiments and Emotions were found sentence-level as well as word-level for entities and keywords.

2 Approach

Our approach focused on two main areas: (1) theme extraction and sentiment analysis and (2) link creation.

The theme extraction step began with a set of news articles marked by hand with their themes and sentiments. For each of these articles, sentence-level analysis data was extracted using IBM's Alchemy API. Then the sentence data was aggregated into a single dataset for each article using a weighted average. The descriptions of the extracted data can be seen in Table 1. A sample of extracted data from a sentence in our dataset has also been included in the Appendix.

During aggregation, article sentiment and emotion were found by performing a weighted average of the sentiment and emotion levels extracted for each sentence and for each entity and keyword. Averages were weighted using relevance scores.

In order to form links between articles, a set of comparison functions was created to determine the relationship between two articles. First, the relationship between the themes of articles was found by comparing the articles' entities, keywords, and taxonomies. A weighted average was performed, calculating the number of matched entities and keywords compared to the total number of entities and keywords found in both articles. Upon tweaking the weight used in this average, the best results were achieved using 1+relevance. This method of weight calculation was similarly applied for taxonomies. If two articles were found to have similar themes, a link was created, then their sentimental and emotional relationships were evaluated.

The relationship between the viewpoint of two articles was calculated by aggregating the maximum emotional difference and the difference in sentiments. If two articles' viewpoints were sufficiently different, the link between the articles was labeled as containing "opposing" viewpoints. On the other hand, if the differences between these articles was insufficient, the link was labeled as "similar".

When the link creation was complete, a visualization of the article relationships was created in the form of a graph. Each article is a node containing its keywords, entities, and taxonomy; each edge represents a thematic relationship and is correspondingly labeled as either "similar" or "opposing" based on the viewpoint relationship aggregation.

	Entites/Keywords	Taxonomy	Sentiment	Emotion
Threshold	0.06	0.25	0.12	0.12
Minimum	0.0	0.0	0.04	0.05
Maximum	0.09	0.31	0.24	0.22

Table 2: Final threshold values for each comparison during link creation and the minimum and maximum observed values. A link is created between articles that meet the entities/keywords and taxonomy comparison thresholds. Articles meeting the sentiment and emotion thresholds are considered to have opposing viewpoints.

3 Experiments and Results

The majority of the experimentation performed during this project was to determine the weights and thresholds used during aggregation steps. First, we found the proper weighting scheme for aggregating sentiment and emotion by evaluating the F_1-score of multiple weights. Out of a combination of baseline, frequency, and relevance weights, we found that relevance gave the best results. Similarly, we evaluated the proper threshold to be used for each of the comparisons: (1) entities and keywords, (2) taxonomy, (3) sentiment, and (4) emotion. For conciseness, we have only included the final results for all thresholds in Table 2.

Upon completion of our project, we have both a file format and a visualization of the linked articles. The visualization is a graph representation of the links between articles. Nodes that contain links are clustered together. Nodes and edges can be expanded to display article and relationship information. Views of the article graph and a sample individual article node from the interactive visualization can be found in Figure 1 and Figure 2 in the appendix. Note that article viewpoints are displayed on the edges.

Our file output contains a list of all linked articles along with their entity, taxonomy, sentiment, and emotion comparison scores and a viewpoint indicator. An example article link is shown below:

```
article 1: GovChristieLeavesGunControlsBehindinNewJersey
article 2: RepublicanCandidatesDeeplyEntrenchedAgainstGunControls
entity:    0.0859298046379
taxonomy:  0.312259510317
sentiment: 0.167339384365
emotion:   0.117104697335
viewpoint: similar
```

4 Discussion

We have reached our baseline goal to develop both:

- A visual representation of article relationships, in the form of a graph with articles as nodes and edges representing relations between articles.
- A file-based representation of the relationships between articles, which can be easily imported and parsed by another program.

We did not meet our stretch goals to extend our module further. We wished to extract the organizations and people mentioned in an article and determine each entity's viewpoint on the article topics. Thus given a set of articles, we can determine the general viewpoints of any particular organization mentioned across the entire set.

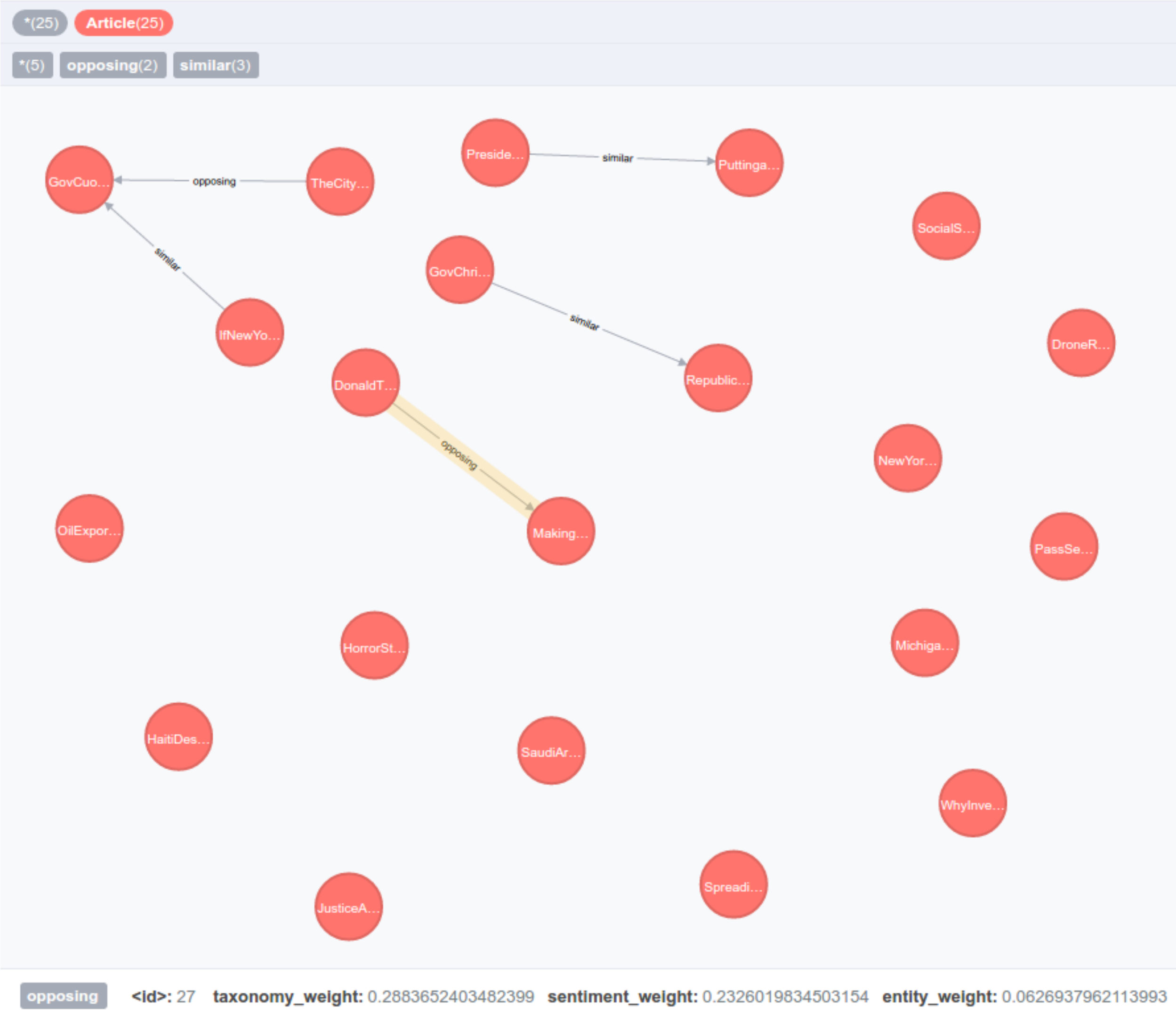
In the future, we will continue to develop our module in order to improve our link quality and incorporate our stretch goals. While these features may not be used by the Opinionated group, we feel they can be utilized in developing more robust comparisons between different organizations as seen by the media. For example this viewpoint extraction could be useful in determining the views of a political party on the topic of building an oil pipeline, or a similar development. We also wish to develop a more advanced graph visualization technique to display article link data in the graph itself. One potential feature within this task is to have links that change the edge width based on the strength of a relationship. Another avenue we wish to pursue in graph visualization is experimenting with the effect of clustering algorithms on the graph to better view larger clusters. One difficulty we may encounter in this experimentation is acquiring enough data to construct a relevant graph. Lastly, our testing consisted of only 25 articles and so in the future we wish to add large scale testing. Similarly to clustering, this testing may prove difficult because it would require a large dataset of articles manually analyzed for theme and viewpoint.

Our group consisted of two members, Courtney Fiala and Anders Maraviglia. For the most part, Courtney contributed to the development of the aggregation methods and the comparison functions. She also wrote the rough drafts of the project proposal and the final report. Anders contributed mainly to the functionality for data extraction and graph visualization. He also created rough draft slides for the proposal and final presentations. Both group members contributed to article parsing, documentation, and proofreading of the project deliverables. A repository of our code can be found at `https://github.com/opinionated/OpinionatedNLPAnalysis`.

5 Appendix

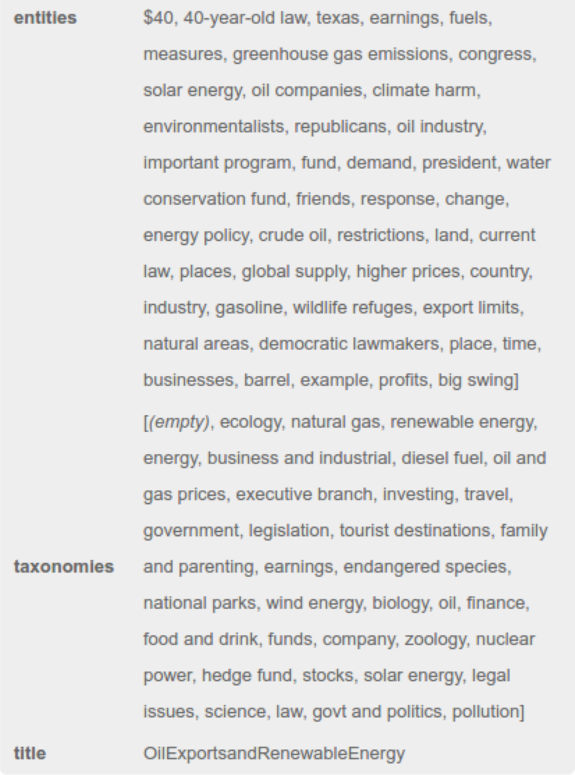
```
docSentiment: negative, -0.87
docEmotions: anger: 0.02, joy: 0.50, fear: 0.04, sadness: 0.44, disgust: 0.03
entities: [
  { text: 10 percent
    sentiment: neutral
    emotions: anger: 0.03, joy: 0.77, fear: 0.07, sadness: 0.15, disgust: 0.02
    relevance: 0.01 } ]
keywords: [
  { text: Chinas main stock
    sentiment: negative, -0.72
    emotions: anger: 0.03, joy: 0.77, fear: 0.07, sadness: 0.15, disgust: 0.02
    relevance: 0.92 }
  { text: commodity prices
    sentiment: negative, -0.84
    emotions: anger: 0.05, joy: 0.14, fear: 0.03, sadness: 0.74, disgust: 0.06
    relevance: 0.59 }
  { text: percent
    sentiment: negative -0.72
    emotions: anger: 0.03, joy: 0.77, fear: 0.07, sadness: 0.15, disgust: 0.02
    relevance: 0.32 } ]
taxonomy: [
  { label: /finance/investing/stocks
    relevance: 0.55 }
  { label: /finance/investing/options
    relevance: 0.53 }
  { label: /business and industrial/agriculture and forestry/crops and seed
    relevance: 0.26 } ]
```

Sample Alchemy API extraction for the sentence "China's main stock index fell nearly 10 percent for the week, depressing stock and commodity prices elsewhere."



<id>: 27 **taxonomy_weight:** 0.2883652403482399 **sentiment_weight:** 0.2326019834503154 **entity_weight:** 0.0626937962113993

Figure 1: A screenshot of the interactive visualization. Out of 25 test articles, 2 were found to pose opposing viewpoints to the same theme and 3 have similar viewpoints for a theme. Note that the label and weights displayed along the bottom of the image correspond to the highlighted edge.



entities	$40, 40-year-old law, texas, earnings, fuels, measures, greenhouse gas emissions, congress, solar energy, oil companies, climate harm, environmentalists, republicans, oil industry, important program, fund, demand, president, water conservation fund, friends, response, change, energy policy, crude oil, restrictions, land, current law, places, global supply, higher prices, country, industry, gasoline, wildlife refuges, export limits, natural areas, democratic lawmakers, place, time, businesses, barrel, example, profits, big swing]
taxonomies	[*(empty)*, ecology, natural gas, renewable energy, energy, business and industrial, diesel fuel, oil and gas prices, executive branch, investing, travel, government, legislation, tourist destinations, family and parenting, earnings, endangered species, national parks, wind energy, biology, oil, finance, food and drink, funds, company, zoology, nuclear power, hedge fund, stocks, solar energy, legal issues, science, law, govt and politics, pollution]
title	OilExportsandRenewableEnergy

Figure 2: A screenshot of node data within the interactive visualization. When a node is clicked, its data can be viewed as a list containing the article title, entities, keywords, taxonomy, sentiment, and emotions.