Clustering of Articles by Theme and Sentiment

Natural Language Processing Term Project Proposal

Courtney Fiala and Anders Maraviglia

October 5, 2016

1 Executive Summary

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 topic 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. This can be applied to display a side-by-side comparison of two articles presenting opposing viewpoints on a specific topic.

2 Goal

We have two major objectives for our project:

1. Extract themes from articles
    - Weight each theme by significance and/or prevalence in the article
    - Categorize each extracted theme as being positive, negative, or neutral
2. Create links between articles that share common themes
    - Quantify the links by the weight of the shared themes in both articles
    - Categorize the link as being either positive or negative

The uniquely defining feature of this project is the use of themes to both relate and define the relation between news articles. This is beneficial to the users (the Opinionated project) since it gives a powerful and potentially more accurate way to cluster articles than simple keyword extraction.

3 Background and Motivation

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 view 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.

There is a group of RCOS students 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. We will attempt to use an in-depth theme extraction and sentiment analysis approach for the article clustering. We will create a module to be used by the Opinionated team in their application to generate the relationship between article viewpoints.

Further, we wish to extend our module to extract the organizations and people mentioned in an article and determine each entity's viewpoint on the article topics. Then given a set of articles, we can determine the general viewpoints of any particular organization mentioned across the entire set.

4 System Architecture and Approach

The algorithm will have two main parts; theme extraction and link creation.

In the theme extraction step, every article in the input set will be parsed by sentence and have any themes and sentiments associated with them identified using either IBM's Alchemy API or Google's Cloud Natural Language API. Each theme found will be stored and counted, along with a sentiment score of either positive, negative, or neutral. After all sentences go through this process, all the themes found will get sorted based on how many times they appeared, and the overall sentiment for each will be decided based on most common score.

Once all articles have gone through the first step, a graph relating all of them can then be created. Links will be formed between every pair of articles that share common themes, where each link will have two critical attributes. The first is weight, which will be based on the difference between the order of theme rankings, where higher weights come from more shared themes being highly ranked in both articles. The second is sentiment, which will be determined by how many shared themes have the same sentiment value, where if most do then the articles agree, otherwise they do not.

After completion of the second step, a graph based around theme clusters will have been formed with links stating the agreement or disagreement between a pair of articles.

5 Deliverables

Upon completion we will deliver to our customer a module such that when a set of articles is inputted, the customer receives:

- 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.

6 Time Table

The table below summarizes our weekly project schedule.

Week	Date	Task
1	10/13	Decide on outside tools to include (determine dependencies and permissions required)
2	10/20	Decide on system architecture and begin coding theme extraction module
3	10/27	Finish theme extraction module
4	11/04	Begin coding sentiment analysis module
5	11/11	Finish sentiment analysis module and begin connecting modules
6	11/18	Continue connecting modules to cluster set of articles
7	11/25	Finish connecting modules and begin entity extraction module
8	12/02	Finish entity extraction module and begin entity viewpoint extraction module
9	12/06	Finish entity viewpoint module and wrap up final deliverables

Our first key assumption that affects the entire project is that we will be able to acquire the tools necessary to extract the theme and sentiment of a particular sentence in an article. We will use these tools to then summarize the viewpoint of the article. A second key assumption that affects week 9 is that our clustering module will perform well in determining an article's sentiment towards a theme. To be able to place a particular view with an organization or person mentioned in the article, we must first correctly assign the viewpoints of the article. Lastly, we assume we can demonstrate the relationships between articles based on their themes and sentiments in such a way that the analysis is meaningful and usable to our customer (RCOS).