12-12-2018

Story Analyzer – an Application using CoreNLP and D3 Visualizations

Michel Mitri
James Madison University, mitrimx@jmu.edu

Story Analyzer – an Application using CoreNLP and D3 Visualizations

Prototype Demonstration

Michel Mitri
James Madison University
mitrimx@jmu.edu

Abstract

Story Analyzer is a software application that helps users visualize and understand a story through the use of natural language processing (NLP) and data visualization APIs. Specifically, Story Analyzer uses Stanford's CoreNLP software library for performing information extraction on a story, and uses D3's JavaScript visualization API to generate a dashboard of results, much which is shown in a set of interrelated and interactive word clouds. The term "story" refers to a narrative that involves people, groups, organizations, or other entities (subjects) performing actions that can affect other people, organizations, or entities (objects). These events occur in certain places and at certain times, and they may include other contextual features of interest. The term "story analysis" pertains to identifying these key elements of the story (subject, objects, actions, time, place, and other contexts), and moreover to represent the relationships between these elements for each event that takes place in the story. The software described in this paper attempts to visually and interactively answer this question: Who did what to whom, where and when did it happen, and what else was going on at the time?

Keywords

Natural language processing, information extraction, word clouds, data visualization.

Description of Story Analyzer

Story Analyzer's system architecture is shown in Figure 1.

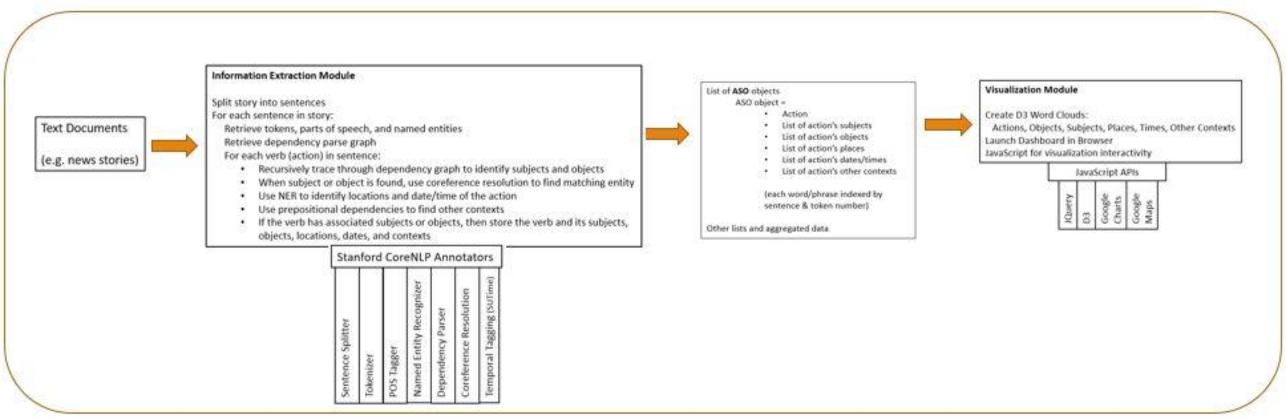

Figure 1. Story Analyzer's system architecture

The application includes contains two main components: information extraction and visualization. The information extraction (IE) algorithm makes extensive use of CoreNLP, as outlined in Figure 1.

Stanford's CoreNLP (Manning et al, 2014) is a Java API consisting of a suite of tools called annotators that can perform many NLP functions. The annotators provide several useful services, including the following:

- Breaking a text document into individual sentences (sentence splitting)

- Tokenizing a sentence (breaking it into individual "words")
- Identifying parts of speech (POS) within a sentence (nouns, verbs, adjectives, adverbs, etc.).
- Named entity recognition – recognizing names of people, places, organizations, dates/times, etc.
- Constituency parsing – constructing taxonomies of noun phrases and verb phrases of a sentence
- Dependency parsing – constructing the graph of dependency relationships between terms in a sentence
- Co-reference resolution – finding all expressions that refer to the same entity in a text
- Temporal tagging – recognizing and normalizing temporal expressions (e.g. "next Wednesday", "the previous summer")

Sentence splitting, tokenizing, and POS tagging are straightforward tasks and CoreNLP is very accurate with these functions. Named entity recognition and parsing are highly useful but are considerably less accurate in NLP's current state of the art, with accuracy rates in the 80%+ range. Coreference resolution, which is vital for connecting related themes in a story, is far less accurate. 60% accuracy is considered a good score in the current NLP climate. So, CoreNLP does not always generate correct results. Nonetheless, the output from these annotators give very useful information for assisting people quickly understand a story narrative.

The main output from the IE module is a list of Action/Subject/Object (ASO) instances. Each ASO represents an action in the story, and includes lists of subjects, objects, places, times, and other contexts associated with that action. These and other data outputs from IE form the input data for the visualizations in the resulting dashboard, shown in Figure 2.

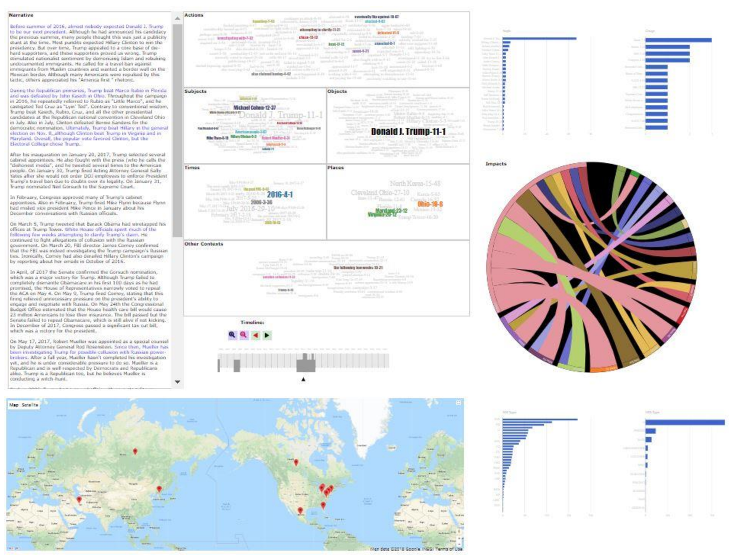

Figure 2. Story Analyzer dashboard

Central to the dashboard are six word clouds depicting the relationships of the ASO instances. These are highly interactive, thanks to D3 (Bostock et al., 2011), a JavaScript API based on scalable vector graphics. When a user hovers over an item in one cloud, the related items in the other clouds are highlighted and grouped according to color. In Figure 2, Donald Trump is selected from the objects cloud. All the actions done to Trump are highlighted, along with their associated subjects, places, times and other contexts. The highlighted sentences in the narrative at the left are the sentences in which Trump is an object of actions.

Story Analyzer's dashboard includes D3 chord and timeline visualizations. The outer band of the chord visualization contains people and groups who are actors (subjects or objects). The chords between any two band segments depict actions that take place between these actors. Selecting a band segment or chord highlights the related items in other visualizations. The timeline depicts a linear time sequence of events, even if the narrative jumps back and forth through time. Picking a timeline segment highlights related items in other visualizations.

The dashboard also includes interactive Google bar charts, maps, and geocoding services for visual depictions of key story elements. Each of these interactively helps users navigate through a story.

This software, and others utilizing NLP and visualization technologies, shows promise for increasing users' efficiency and depth of understanding when reading text.

Story Analyzer's IE and visualization modules are designed specifically for narrative text. It is well-suited for news articles, histories, fictions, etc. It is not as well suited for representing technical documentation, research articles, medical articles, etc.

Possible users of this tool include anyone who needs to do a lot of narrative text reading in their jobs. This requirement is common to many professions, such as consulting, education, intelligence analysis, journalism, and others.

References

Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S.J., and McClosky, D. 2014. "The Stanford CoreNLP Natural Language Processing Toolkit," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, Baltimore, MD: June 22-27, 2014, pp. 55-60.

Bostock, M., Ogievetsky, V., and Heer, J. 2011, "D3 data-driven documents" *IEEE transactions on visualization and computer graphics* (17, 12), pp. 2301-2309.