Intelligent Text Extraction and Summarization for an Improved Community Initiative

Xingbang Liu and Dr. Janyl Jumadinova
Department of Computer Science, Allegheny College
Meadville, PA

ALLEGHENY COLLEGE

https://www.cs.allegheny.edu
jjumadinova@allegheny.edu
liux2@allegheny.edu

Project Summary

We present an intelligent system that was used to process textual information, generate knowledge, and automatically summarize key findings of the My Meadville community statements.

- My Meadville is a non-profit organization with the goal of highlighting the positive work that is being done in the city of Meadville, PA, and bringing upfront the enhancements and improvements that can be made.
- My Meadville conducted a large number of interviews with the residents of Meadville during the community events and transcribed these interviews into textual data files.
- **Our system processes these community statements, finds important keywords, and then produces a summary of the key excerpts from all data.**

My Meadville Manual Data Analysis

Value Statement: Health and Safety
We will be a healthy community in which everyone has access to health care, fresh food, quality housing, and support services.

Supporting Data: Fresh food

- "We value the different variety of foods available in Meadville."
- " Value a community that puts a focal point on healthy food options grown locally."
- "... Values the family businesses and small food places ..."

Figure: My Meadville Organization

System Design

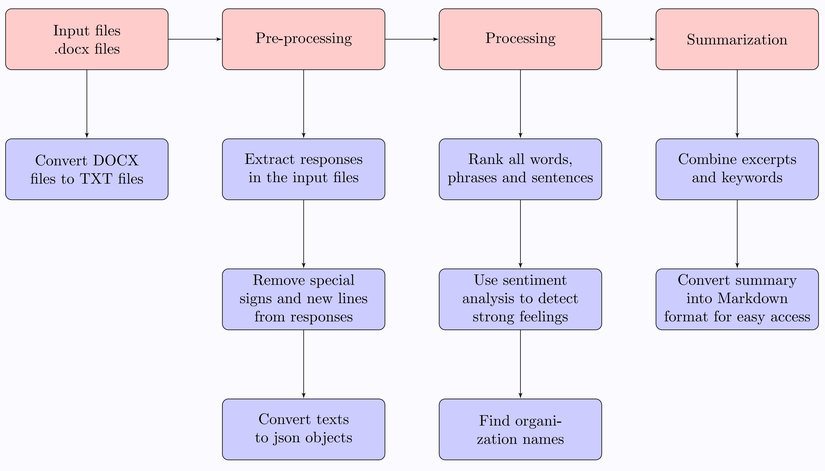

Figure: Different portions comprising the system

Methodology: Key Points

Processing (three layers):

1. The **text rank** layer builds a word graph for voting on the importance based on [Mihalcea 2004].
2. For **sentiment analysis** Scikit-learn was used to implement naive Bayes variations, Logistic Regression, Linear support vector clustering, and stochastic gradient descent classifier algorithms.
3. Stanford **Named Entity Recognizer** was used to locate organization names as they are all important.

Example:

First Layer:

```
{"graf": [[0, "Uh", "uh", "UH", 0, 0],
  [1, "more", "more", "JJR", 1, 1],
  [2, "children", "child", "NNS", 1, 2],
  [3, "activities", "activity", "NNS", 1, 3],
  [0, "and", "and", "CC", 0, 4],
  [4, "events", "event", "NNS", 1, 5],
```

Second Layer:

```
{"count": 1, "ids": [1, 2, 3], "pos": "np",
  "rank": 0.08034322728390975,
  "text": "more children activities"}
{"count": 1, "ids": [15, 16], "pos": "np",
  "rank": 0.06691920745000741, "text": "hefty fee"}
```

Third Layer:

```
{"dist": 0.08593524452032889, "idx": 0,
  "text": "Uh more children activities and events ."}
{"dist": 0.05285600431403769, "idx": 4,
    "text": "but it all comes with a hefty fee too ."}
```

Transcribed Data

Over 200 interviews	Average interview 9,161 words
Min words = 334	Max words = 30,465

Sample Outcome

- Text summary and keywords were extracted from each interview.

Uh more children activities and events . But it would be nice if there was more opportunities for them . Um , there 's so many opportunities for dance and gymnastics and that stuff but it all comes with a hefty fee too .
Keywords: more opportunities, hefty fee, more children activities, stuff

I think those are nice events , and the downtown mall. they did a few years ago
Keywords: nice events, downtown mall, friendly activities, a few years ago

Conclusion and Future Work

- Our findings are being used by My Meadville to create community value statements, highlight relevant community assets and to develop an action plan based on the concerns and areas of improvement identified by the community members.
- In the future we plan to extend our learning framework with better training data and more customized algorithms for sentiment analysis.