Unstructured Text Documents

Twitter/Social Media collections are many unstructured text documents. Unstructured text documents are hard to analyze. Many authors, redundant information. Can accumulate many of these documents in short time.

Summarizing Unstructured Documents

Could extract common information and present a word cloud. Word clouds good at a glance to gain overarching theme. Word clouds lose concepts and structure. How do we maintain semantic representation?

SentenTree

Node-link visualization with force-directed placement. Edge between words indicates occurrence in same tweet. Spatial arrangement is syntactic ordering. Large font indicates high frequency of occurrence.

Frequent Sequential Patterns

Initialization steps:

- Normalize tweets
- Perform tokenization
- Root node of tree of sequential patterns is initial pattern
- Initial pattern contains no words
- Grow new sequential patterns from the root

Interaction Demo

https://twitter.github.io/SentenTree/
Visual Encoding

SentenTree uses a constrained force-directed placement algorithm

Placement constraints: word order, vertical, horizontal

Considerations: Tokenization

Stop words and punctuation removed
Numbers, hashtags, urls, @ handles are matched
No stemming performed

Critique

The Bad:
No stemmer
Final visualizations are still sometimes ambiguous

The Good:
System accomplishes design goals
Well written paper, easy to understand examples
Scalable

Thanks!
Questions?