Visualizing Social Media Content with SentenTree

Mengdie Hu, Krist Wongsuphasawat, John Stasko. IEEE TVCG 23(1):621-630 2017 (Proc. InfoVis 2016)

Presented by: David Johnson

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

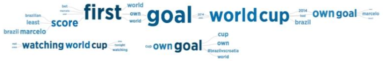
Could extract common information & present a world cloud

Word clouds good at a glance to gain overarching theme

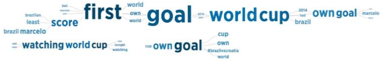
World clouds lose concepts and structure

How do we maintain semantic representation?

SentenTree



SentenTree



Node-link visualization with force-directed placement

Edge between words indicates occurrence in same tweet

Spatial arrangement is syntactic ordering

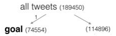
Large font indicates high frequency of occurrence

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

Frequent Sequential Patterns

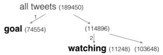


all tweets (189450)

1

goal (74554) (114896)

Frequent Sequential Patterns



all tweets (189450)

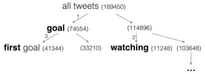
1

goal (74554) (114896)

2

watching (11248) (103648)

Frequent Sequential Patterns



all tweets (189450)

1

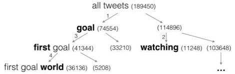
goal (74554) (114896)

3 2

first goal (41344) (33210) **watching** (11248) (103648)

...

Frequent Sequential Patterns



all tweets (189450)

1

goal (74554) (114896)

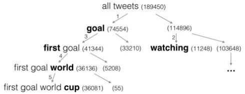
3 2

first goal (41344) (33210) **watching** (11248) (103648)

4

first goal **world** (36136) (5208) ...

Frequent Sequential Patterns



all tweets (189450)

1

goal (74554) (114896)

3 2

first goal (41344) (33210) **watching** (11248) (103648)

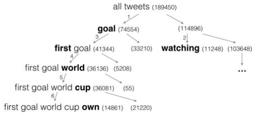
4

first goal **world** (36136) (5208) ...

5

first goal world **cup** (36081) (55)

Frequent Sequential Patterns



all tweets (189450)

1

goal (74554) (114896)

3 2

first goal (41344) (33210) **watching** (11248) (103648)

4

first goal **world** (36136) (5208) ...

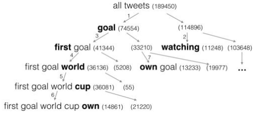
5

first goal world **cup** (36081) (55)

6

first goal world cup **own** (14861) (21220)

Frequent Sequential Patterns



all tweets (189450)

1

goal (74554) (114896)

3 2

first goal (41344) (33210) **watching** (11248) (103648)

4 7

first goal **world** (36136) (5208) **own** goal (13233) (19977) ...

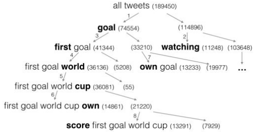
5

first goal world **cup** (36081) (55)

6

first goal world cup **own** (14861) (21220)

Frequent Sequential Patterns



all tweets (189450)

1

goal (74554) (114896)

3 2

first goal (41344) (33210) **watching** (11248) (103648)

4 7

first goal **world** (36136) (5208) **own** goal (13233) (19977) ...

5

first goal world **cup** (36081) (55)

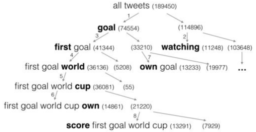
6

first goal world cup **own** (14861) (21220)

8

score first goal world cup (13291) (7929)

Frequent Sequential Patterns



all tweets (189450)

1

goal (74554) (114896)

3 2

first goal (41344) (33210) **watching** (11248) (103648)

4 7

first goal **world** (36136) (5208) **own** goal (13233) (19977) ...

5

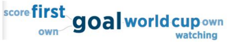
first goal world **cup** (36081) (55)

6

first goal world cup **own** (14861) (21220)

8

score first goal world cup (13291) (7929)



Interaction Demo

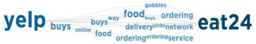
https://twitter.github.io/SentenTree/

Visual Encoding

SentenTree uses a constrained force-directed placement algorithm

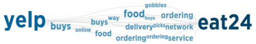
Placement constraints: word order, vertical, horizontal

Visual Encoding

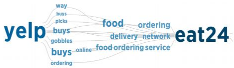


Only word order constraint applied

Visual Encoding



Only word order constraint applied



Horizontal and vertical constraints added

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

Critique

The Good:

System accomplishes design goals

Well written paper, easy to understand examples

Scalable

Thanks!

Questions?