Machine Reading of Natural Language and Interactive Visualization

Anoop Sarkar
Simon Fraser University

© Matt Lum

SFU NatLangLab

natlang.cs.sfu.ca

Jasneet Sabharwal

Orland Hoeber

Andrei Vacariu

My collaborators on this project

Ramtin Seraj

Maryam Siahbani

Max Whitney

Summarization

SQuASH: SFU QA Summarization System

Input: 25 news articles,

Complex question Output: 250-word summary

Q. Describe developments in the movement for the independence of Quebec from Canada.

A. Canadian Prime Minister Jean Chretien has ruled out the possibility that the French-speaking province of Quebec could declare independence even if a majority of Quebeckers vote for secession in the next referendum. Premier Lucien Bouchard and the separatist Parti Quebecois have retained control of Canada's largest province, possibly setting the stage for a bitter confrontation with the rest of Canada, and possibly another referendum on independence. Canada's Supreme Court managed to rule on the question of Quebec secession without infuriating leaders on either side of a bitter ideological divide.
...

What's wrong with this picture?

Canadian Prime Minister Jean Chretien has ruled out the possibility that the French-speaking province of Quebec could declare independence even if a majority of Quebeckers vote for secession in the next referendum. Premier Lucien Bouchard and the separatist Parti Quebecois have retained control of Canada's largest province, possibly setting the stage for a bitter confrontation with the rest of Canada, and possibly another referendum on independence. Canada's Supreme Court managed to rule on the question of Quebec secession without infuriating leaders on either side of a bitter ideological divide. ...

- Difficult to evaluate.
- The summary is static. Not interactive.
- Completely *new* summary in response to new questions.
- Many different facets in the question and summary:
  - people, locations, predicates (cannot select facets)
- Can NLP help people create better summaries?

Interactive visualization for language summarization

One problem:

Language is tough to visualize

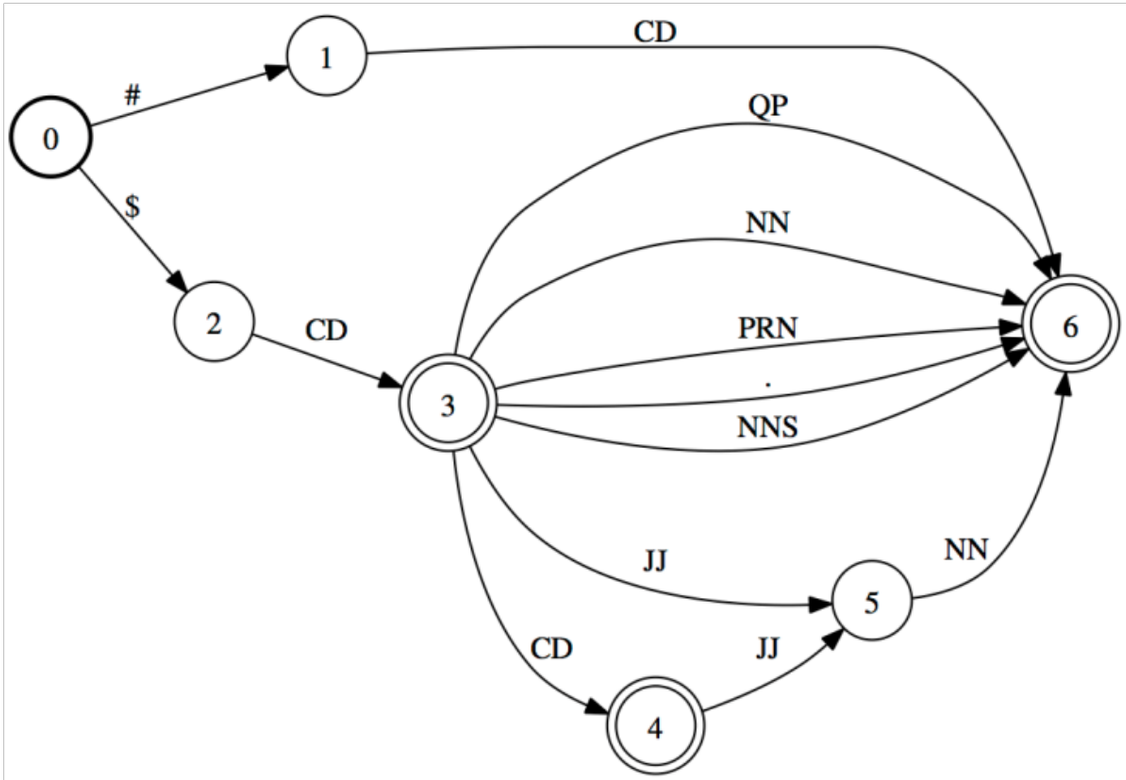
SUBJECT PUNCHED QUICKLY OXIDIZED TCEJBUS DEHCNUP YLKCIUQ DEZIDIXO
CERTAIN QUICKLY PUNCHED METHODS NIATREC YLKCIUQ DEHCNUP SDOHTEM
SCIENCE ENGLISH RECORDS COLUMNS ECNEICS HSILGNE SDROCER SNMULOC
GOVERNS PRECISE EXAMPLE MERCURY SNREVOG ESICERP ELPMAXE YRUCREM
CERTAIN QUICKLY PUNCHED METHODS NIATREC YLKCIUQ DEHCNUP SDOHTEM
GOVERNS PRECISE EXAMPLE MERCURY SNREVOG ESICERP ELPMAXE YRUCREM
SCIENCE ENGLISH RECORDS COLUMNS ECNEICS HSILGNE SDROCER SNMULOC
SUBJECT PUNCHED QUICKLY OXIDIZED TCEJBUS DEHCNUP YLKCIUQ DEZIDIXO
CERTAIN QUICKLY PUNCHED METHODS NIATREC YLKCIUQ DEHCNUP SDOHTEM
SCIENCE ENGLISH RECORDS COLUMNS ECNEICS HSILGNE SDROCER SNMULOC

Language is not pre-attentive: must foveate to process the information

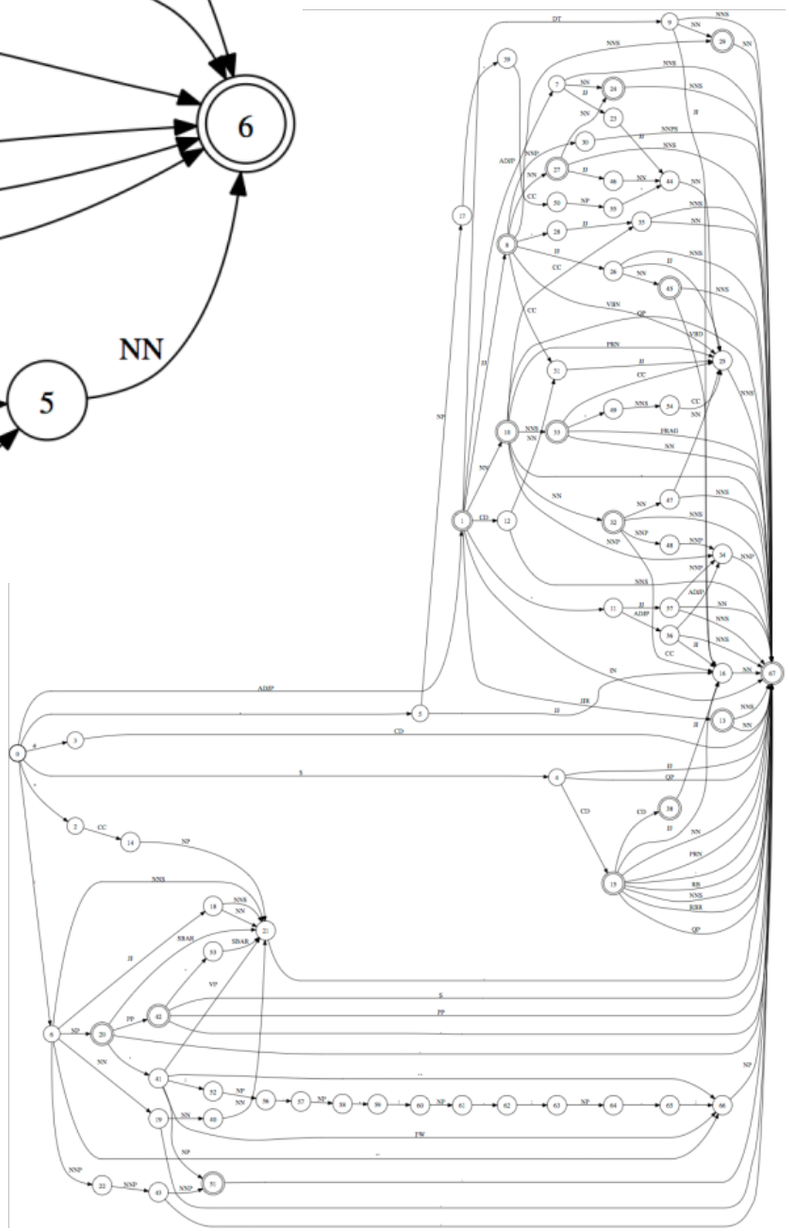
$ 20 happy meal

$ CD JJ NN

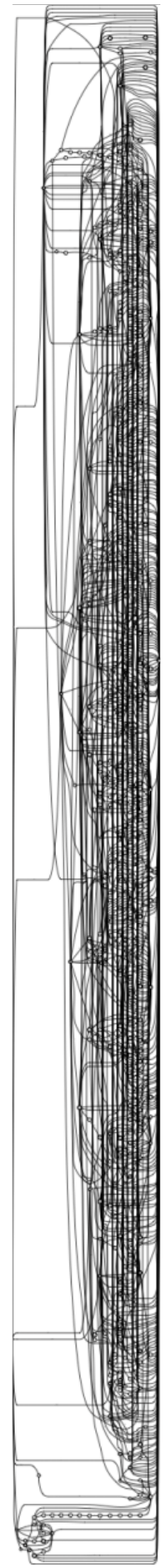
10 noun phrase types



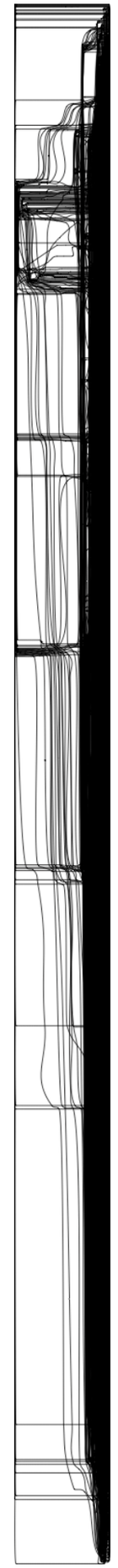
100 noun phrase types



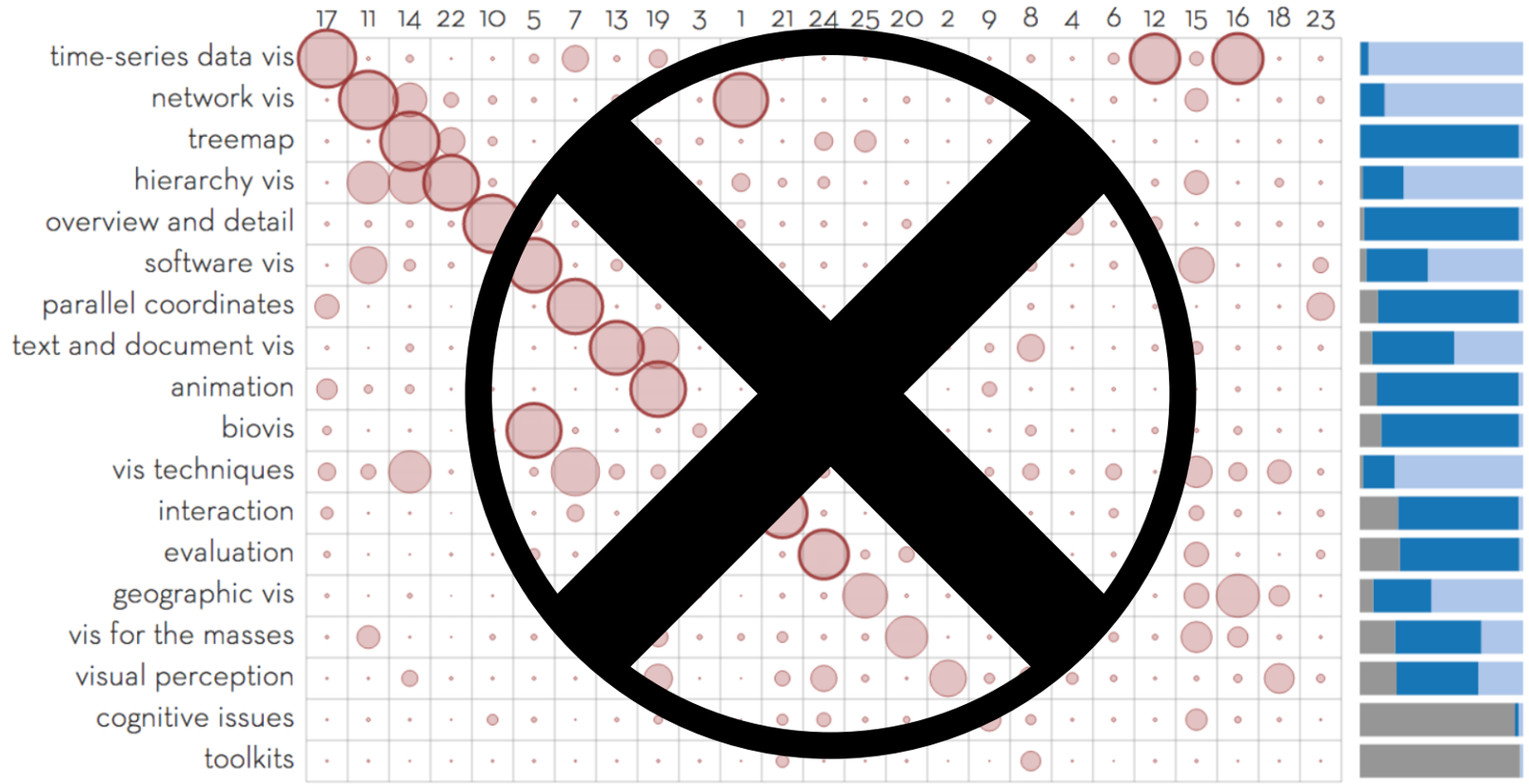
1000 noun phrase types



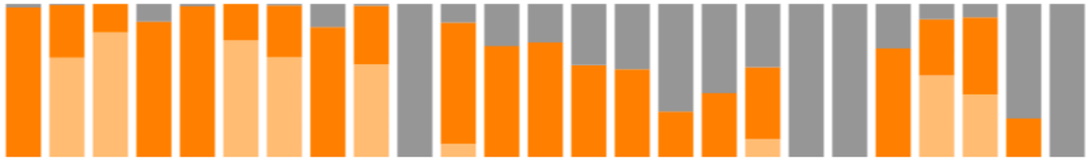
6K noun phrase types



- the engine.



- the engine caught fire.



- Passengers believed the engine caught fire.

- Passengers reported they saw streaks of flames out of the engine and believed the engine had caught fire.

Our Approach

Summarization using Interactive Visualization

- Exploit (shallow) language understanding: who did what to whom, when, where and how.

- Visualize “things, not strings”*: exploit entity-linking of mentions in the text

- Map text spans to spatial, temporal, social views.

- From high dimensional words, phrases to low dimensional visualizations

*http://googleblog.blogspot.ca/2012/05/introducing-knowledge-graph-things-not.html

Lensing Wikipedia

Semantic roles and Entities

Role	Person	Organization	Location
Clear selection	Clear selection	Clear selection	Clear selection
theme [7555]	Augustus [445]	East India Company [15	United States [4746]
agent [7342]	Adolf Hitler [421]	NBC [139]	England [3317]
topic [3282]	Hannibal [370]	CBS [122]	Rome [3092]
patient [3216]	Charlemagne [331]	Boeing [105]	France [2318]
thing set [3138]	Julius Caesar [321]	Metro-Goldwyn-Mayer [	China [2063]
entity changing [2969]	Ptolemy [299]	Dutch East India Compa	Athens [1983]
new state [2958]	Heraclius [283]	American Broadcasting (	Italy [1758]
corpse [2693]	Antiochus III the Great [2	BBC [34]	London [1625]
comment [2366]	Charles Martel [251]	The Walt Disney Compa	Germany [1352]
task [2016]	Alexander the Great [23	Ford Motor Company [3	Egypt [1305]
entity defeated [1753]	Demosthenes [233]	The New York Times [33	Japan [1279]
killer [1719]	Belisarius [199]	Microsoft [33]	Constantinople [1213]
creator [1695]	Pericles [196]	United Airlines [33]	Spain [1196]
entity victorious [1598]	Hadrian [196]	IBM [31]	Sparta [1195]

Person: Napoleon ×

d Entity facets

1879 CE: attach

June 1 – Napoléon, Prince Imperial (Napoléon IV), great-nephew of Napoléon Bonaparte, Bonapartist pretender to the French throne, is killed in Africa while **attached** to the British Army during the Anglo-Zulu War.

1879 CE: kill

June 1 – Napoléon, Prince Imperial (Napoléon IV), great-nephew of Napoléon Bonaparte, Bonapartist pretender to the French throne, is **killed** in Africa while attached to the British Army during the Anglo-Zulu War.

1840 CE: place

December 15 – The corpse of Napoleon is **placed** in the Hôtel des Invalides in Paris.

1840 CE: bury

He is **buried** in Les Invalides. [+]

Role

Clear selection

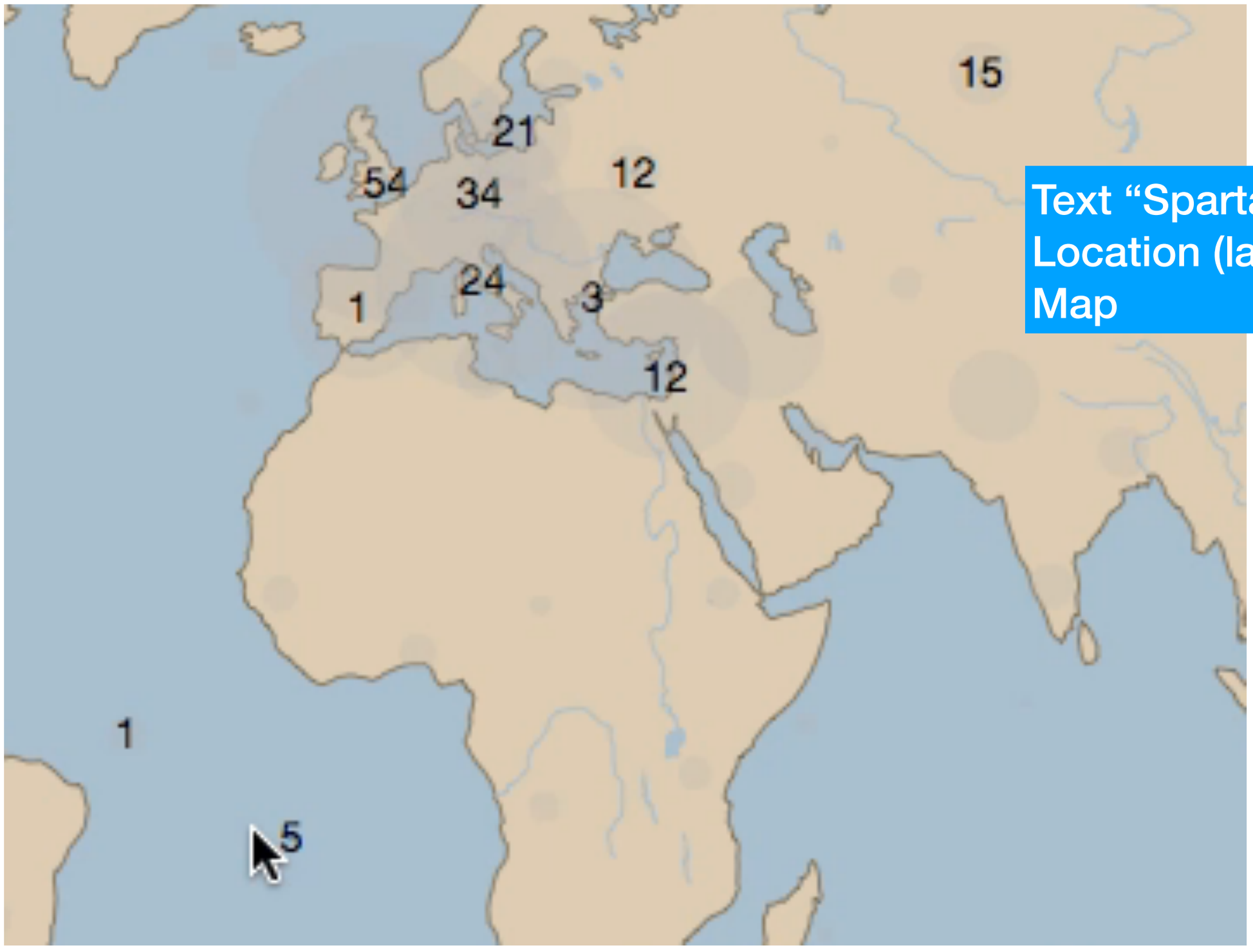
entity defeated [22]
agent [21]
entity victorious [20]
theme [12]
topic [8]
comment [6]
corpse [5]
impelled action [4]
beginner [4]

Person

Clear selection

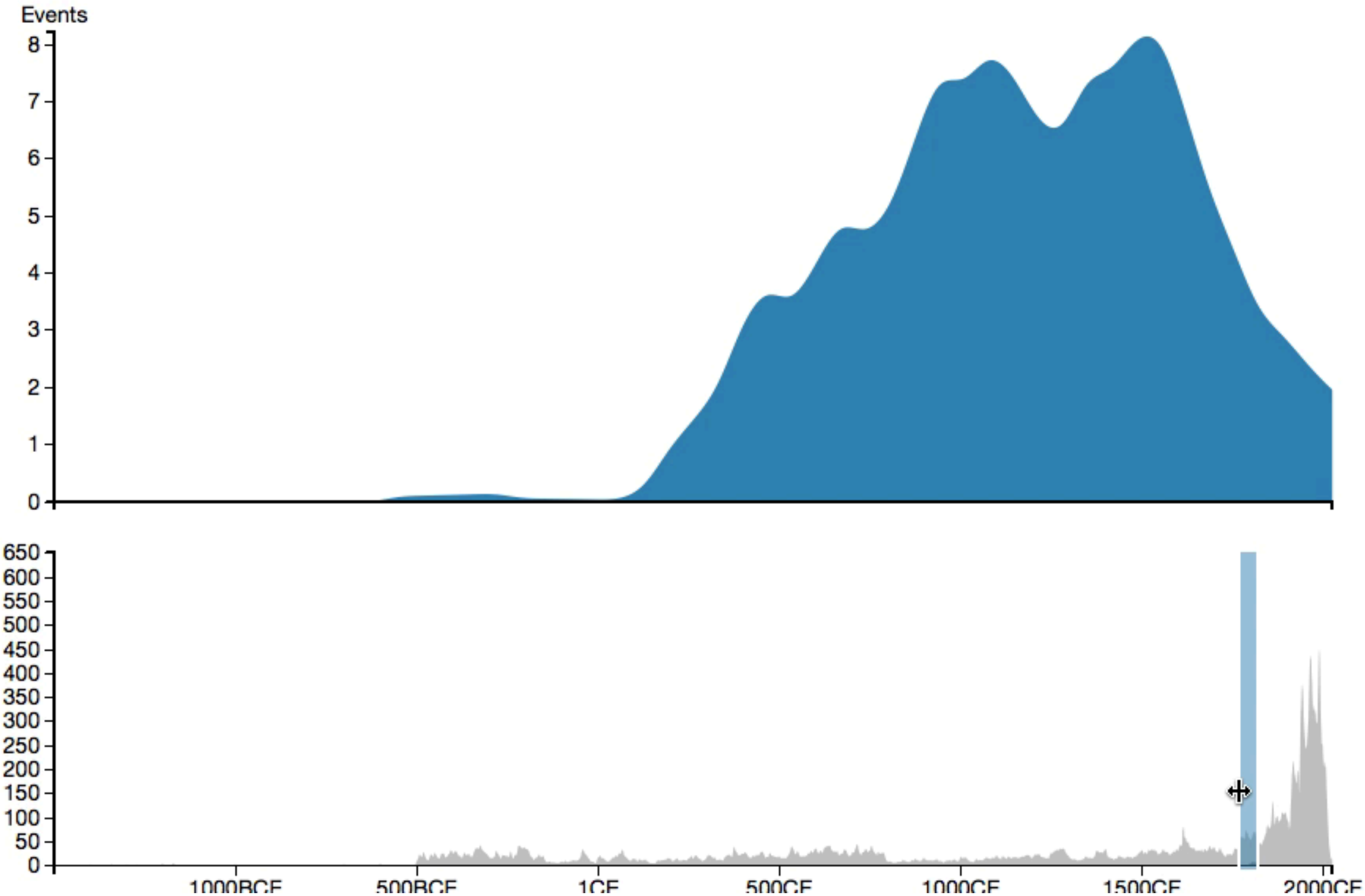
Napoleon [157]
Alexander I of Russia [6]
Levin August, Count von
Louis-Nicolas Davout [4]
Simón Bolívar [3]
Scipione Borghese [3]
Archduke Charles, Duke
Charles William Ferdinar
Karl Mack von Leiberich

Location mentions in text



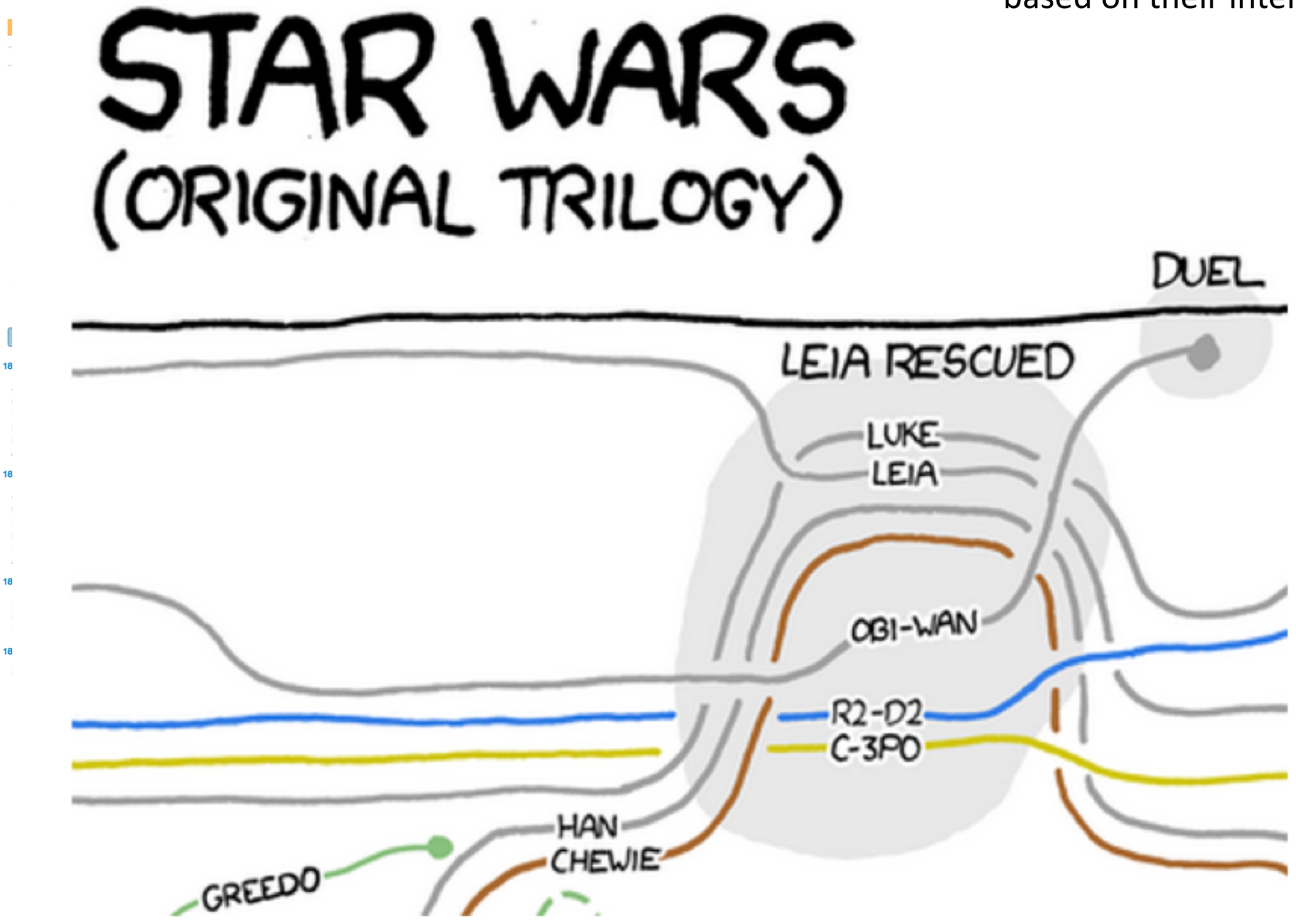
Text “Sparta” :-
Location (lat, long) :-
Map

Timeline of Events

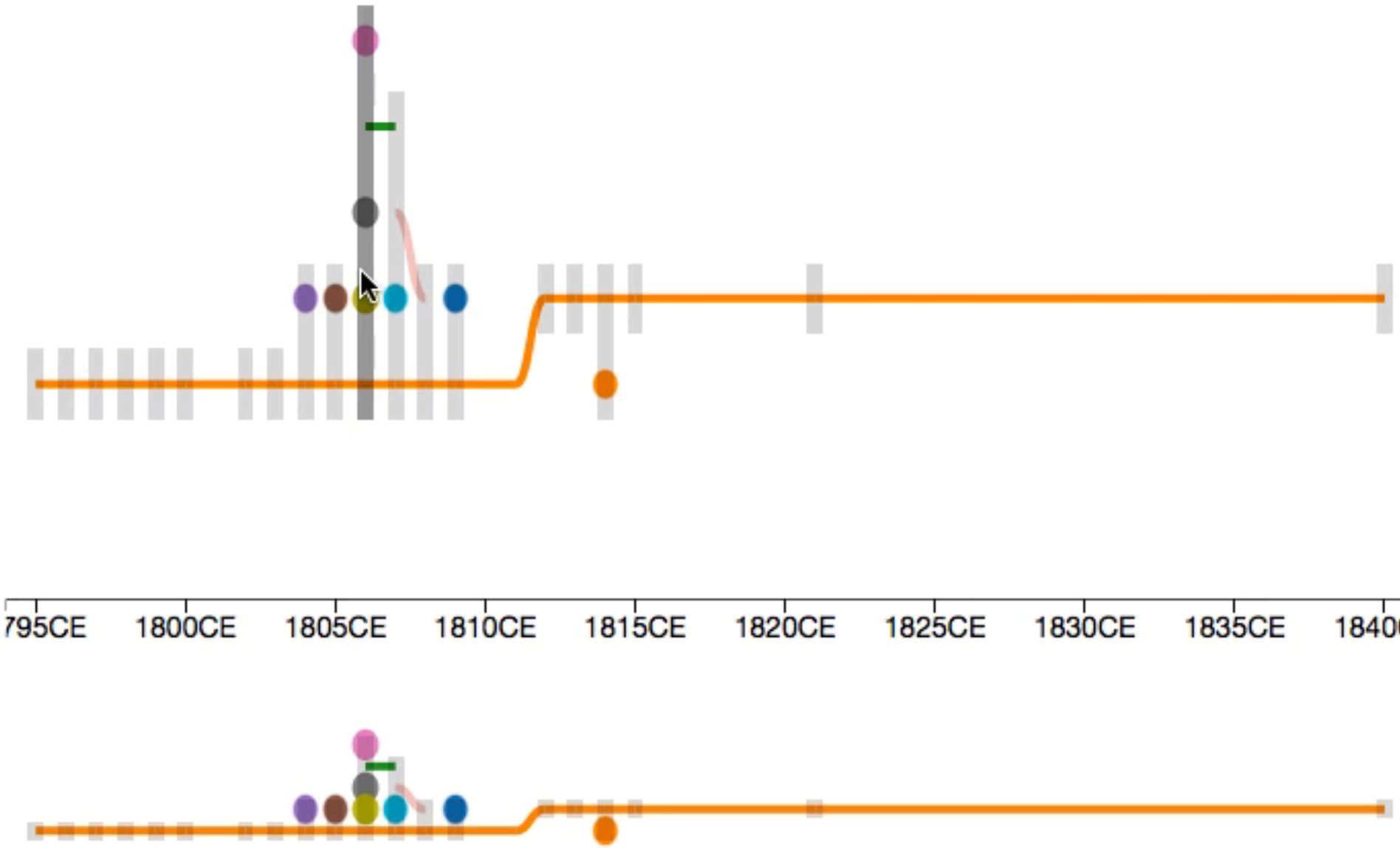


Storylines

Narrative movie chart from http://xkcd.com/657 showing entities arranged in a timeline in clusters based on their interactions



Storylines



Napoleon

Simón Bolívar

Karl Mack von Leiberich

Levin August, Count von Bennigsen

Charles William Ferdinand, Duke of Brunswick-Wolfenbüttel

Frederick Louis, Prince of Hohenlohe-Ingelfingen

Louis-Nicolas Davout

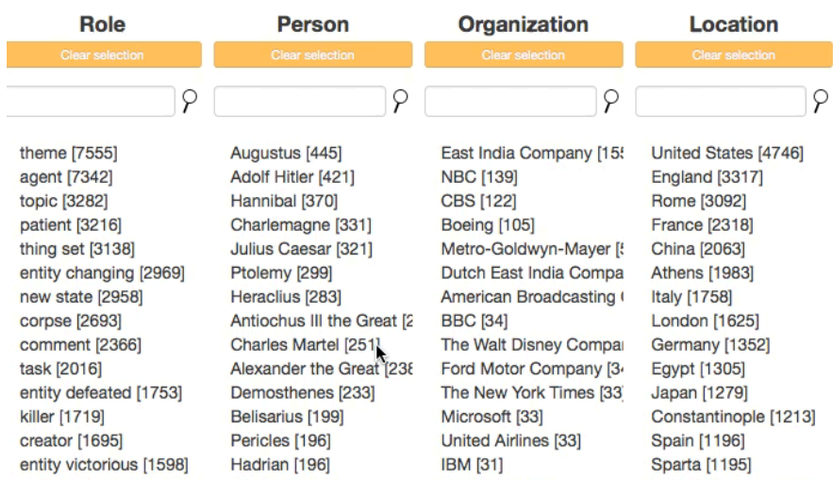
Alexander I of Russia

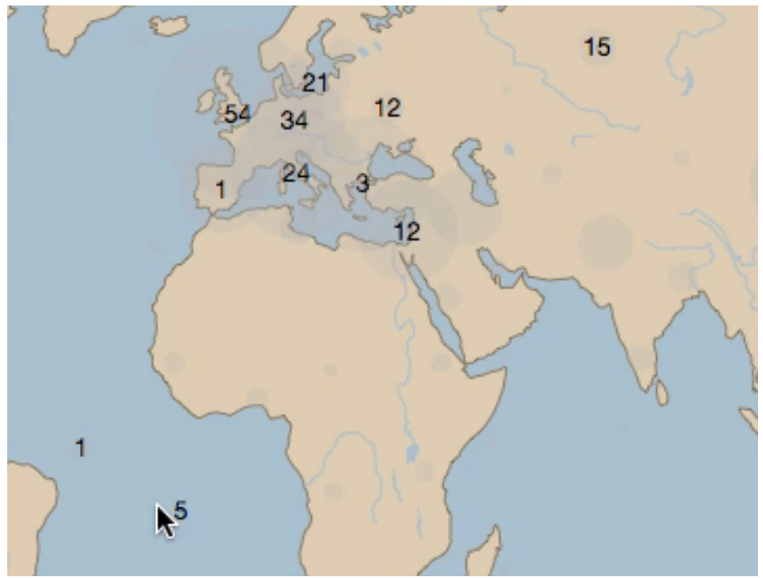
Scipione Borghese

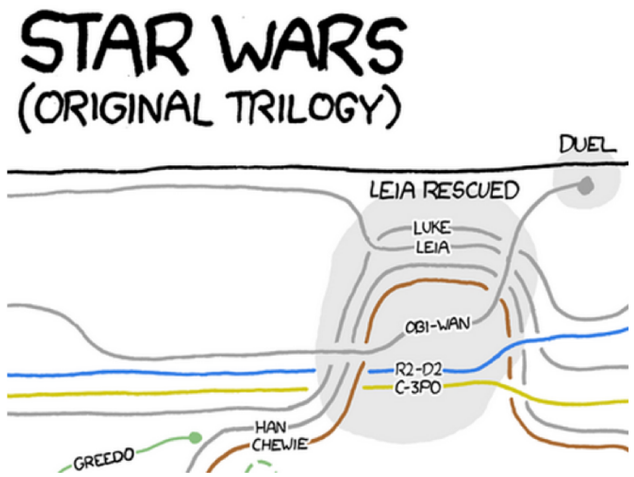
Archduke Charles, Duke of Teschen

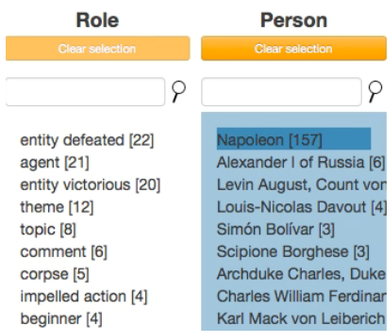
Gebhard Leberecht von Blücher

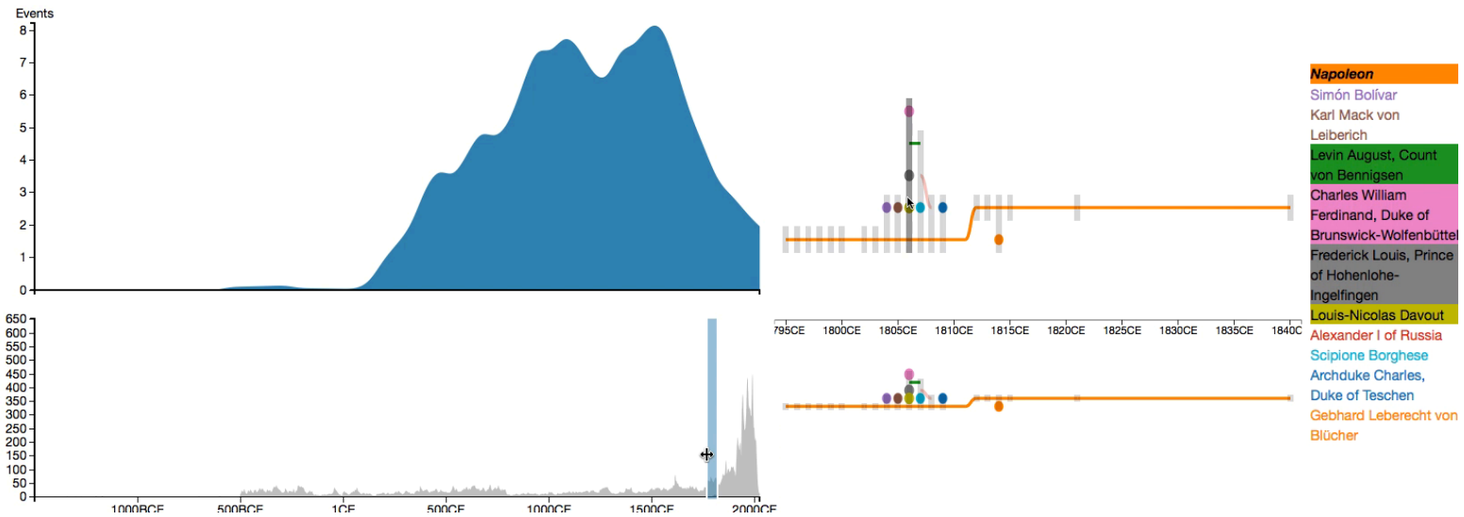
Lensing Wikipedia











Notes

On May 12 1797, Napoleon conquered Venice. The last ruler of Venice, referred to as the Doge, was Ludovico Manin. He was forced to step down as Doge by Napoleon. This event ended 1,100 years of independence for Venice.

From text data to visualization

1471

From Wikipedia, the free encyclopedia

Events [edit]

January–December [edit]

- March 1 – Emperor Le Thanh Tong captures the Champa Capital, establishing new regions in middle Vietnam.
- March – The Yorkist King Edward IV returns to England to reclaim his throne.
- April 14 – Battle of Barnet: Edward defeats the Lancastrian army under Warwick, who is killed.

March 1 – Emperor Le Thanh Tong captures the Champa Capital, establishing new regions in middle Vietnam

Semantic Role Labeling

March 1 – Emperor Le Thanh Tong captures the Champa Capital, establishing new regions in middle Vietnam

predicate	capture
arg0	Emperor Le Thanh Tong
arg1	the Champa Capital

Semantic Role Labeling

- Semantic role labeling is a well studied NLP task
- Can be modelled as structured prediction
- First step: syntactic parsing of the text to obtain mention spans
- Second step: identify and classify semantic roles based on the syntactic structure
- Trained on the PropBank corpus (1M words of Wall Street Journal) — need for domain adaptation

Semantic Role Labeling

March 1 – Emperor Le Thanh Tong captures the Champa Capital, establishing new regions in middle Vietnam

predicate capture

arg0 Emperor Le Thanh Tong

arg1 the Champa Capital

Named Entity Recognition

person Emperor Le Thanh Tong

location Vietnam

Semantic Role Labeling

location Vietnam

Vietnam

From Wikipedia, the free encyclopedia

Coordinates: 16°10′N 107°50′E

Vietnam (UK: /ˌvjɛtˈnæm, -ˈnɑːm/, US: /ˌviːətˈnɑːm, -ˈnæm/ (listen);[10] Vietnamese: *Việt Nam* [vîət nāːm] (listen)), officially the **Socialist Republic of Vietnam** (Vietnamese: *Cộng hòa xã hội chủ nghĩa Việt Nam* (listen)), is the easternmost country on the Indochina Peninsula in Southeast Asia. With an estimated 94.6 million

Socialist Republic of Vietnam
Cộng hòa xã hội chủ nghĩa Việt Nam (Vietnamese)

Flag

Emblem

Semantic Role Labeling

March 1 – Emperor Le Thanh Tong captures the Champa Capital, establishing new regions in middle Vietnam

predicate	capture
arg0	Emperor Le Thanh Tong
arg1	the Champa Capital
location	Vietnam

Semantic Role Labeling

March 1 – Emperor Le Thanh Tong captures the Champa Capital, establishing new regions in middle Vietnam

arg0	Emperor Le Thanh Tong
arg1	the Champa Capital
getter	Emperor Le Thanh Tong
thing gotten	the Champa Capital

Predicate sense disambiguation

Mary gamely kicked in $5 towards John’s bail .

predicate kick

	kick.01
arg0	kicker
arg1	thing kicked
arg2	instrument

	kick.03
arg0	contributor
arg1	contribution
arg2	given to

Predicate sense disambiguation

- Train a classifier to predict the **verbose labels** for each argument of a predicate, “arg0”, “arg1”, etc.

- Exploit metadata written by annotators on the PropBank corpus (1M words of Wall Street Journal).

- Our best classifier obtained 92% accuracy on the PropBank corpus. Performance is lower on Wikipedia.

- Creates a lightweight ontology that is predicate-centric

- Some of the human annotations are a bit strange, e.g. for predicate “resurrect”.

Demo

User Study

Interaction Types

Interaction Type	Lensing Wikipedia
Input	Faceted browsing of entities
Control	Remove constraints
Informational	Storyline, Timeline, Comparison, Map, Follow link to Wikipedia
Personalization	Return to document, Edit notes

The Task

- User study design based on Vakkari (2003)

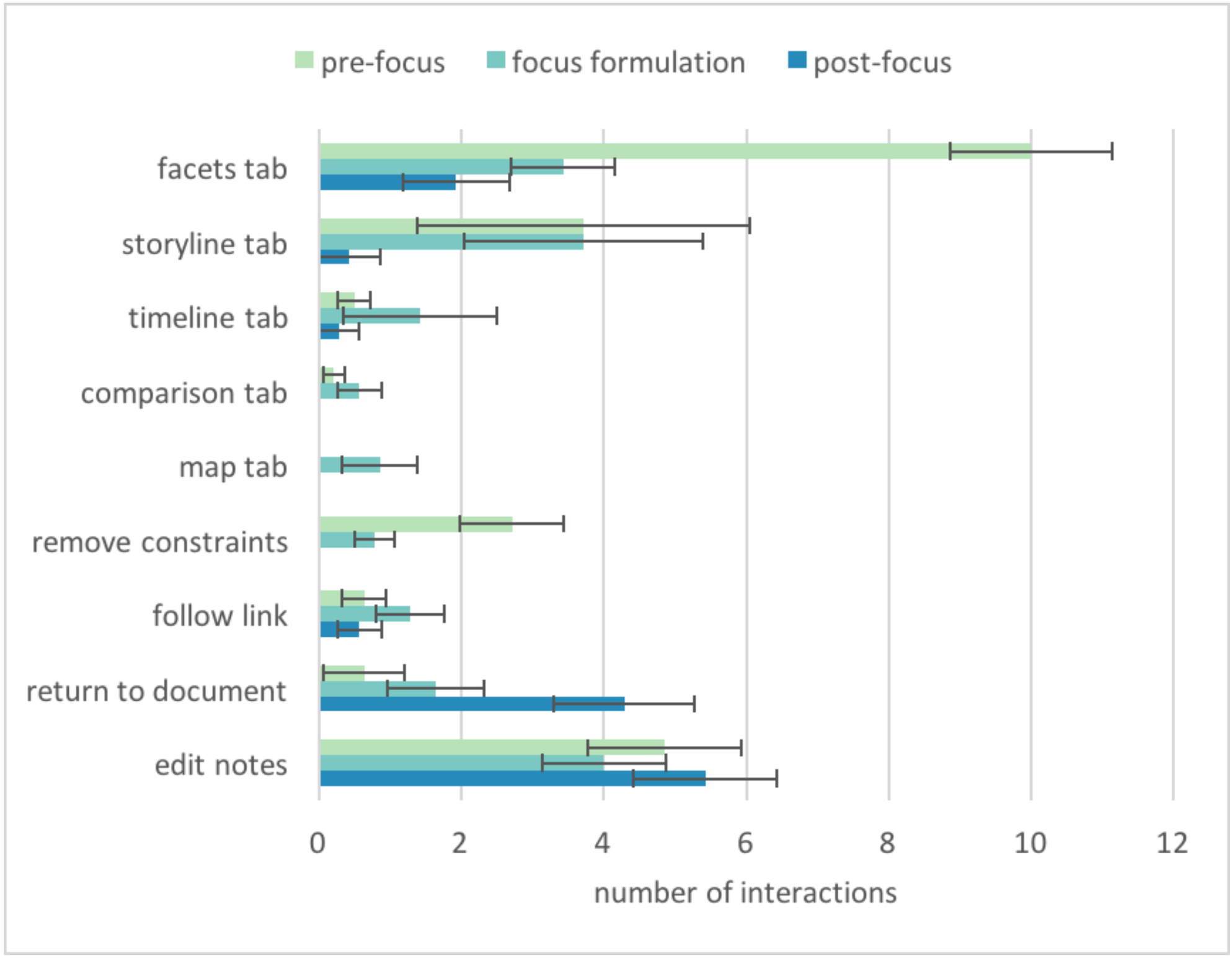
- Write a 200 word outline for an essay on relationship between [primary person] and another person in history.

1. Find a small set of important people related to [primary person]. Pre-focus

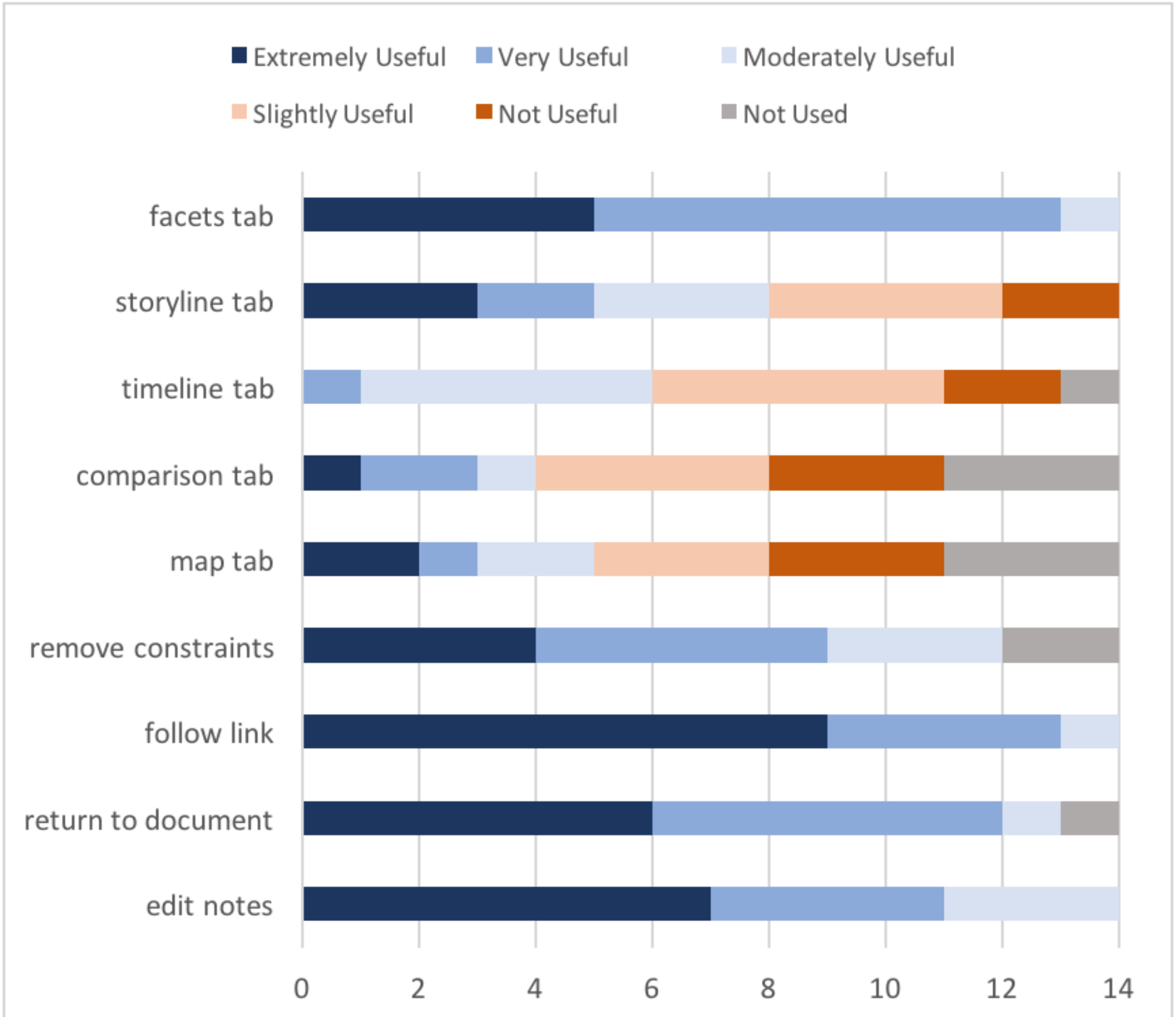
2. Narrow focus down to one person and study the relationship to [primary person]. Focus formulation

3. Collect 5 Wikipedia articles and write the 200 word summary for the essay. Post-focus

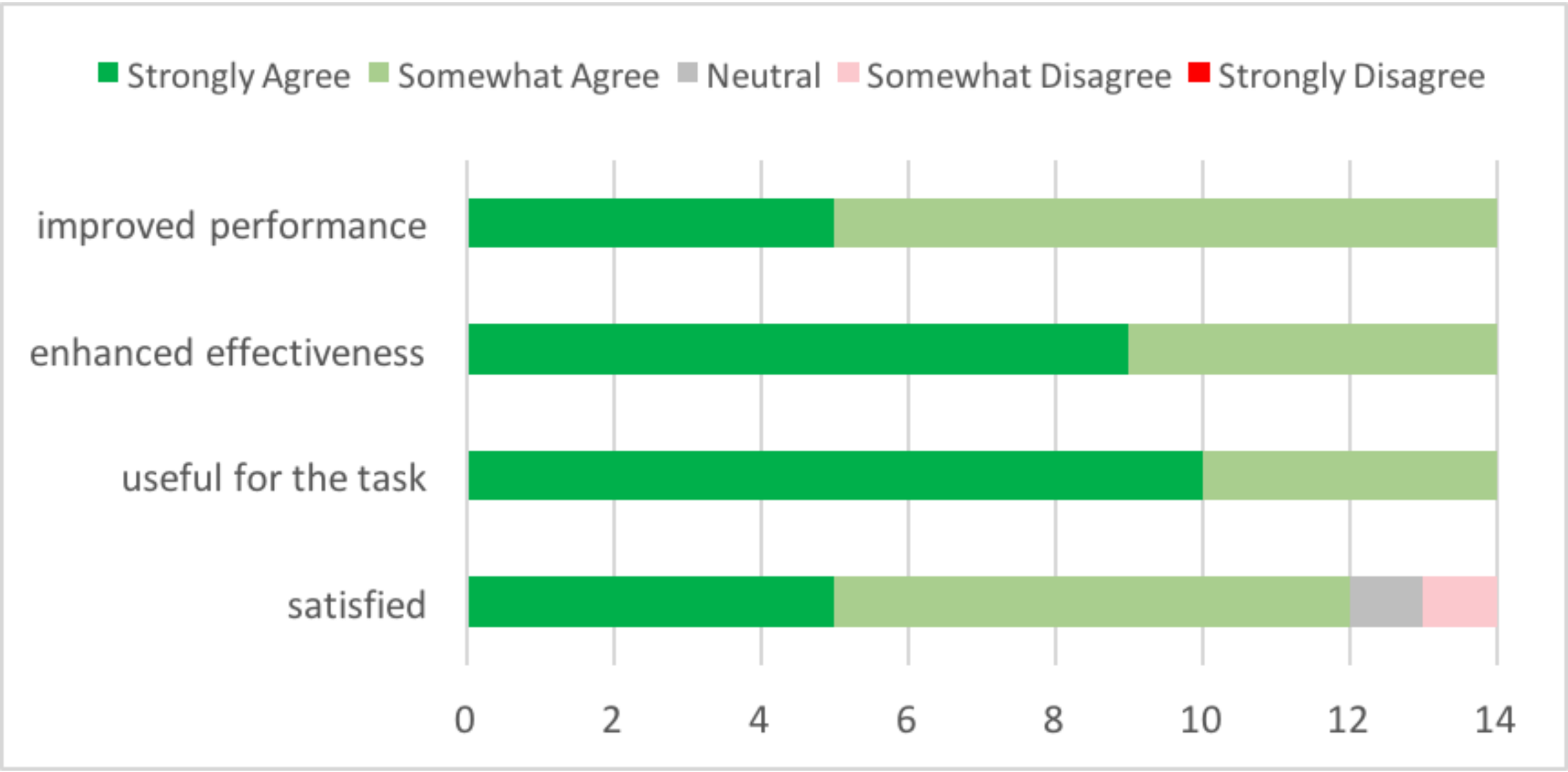
Number of Interactions



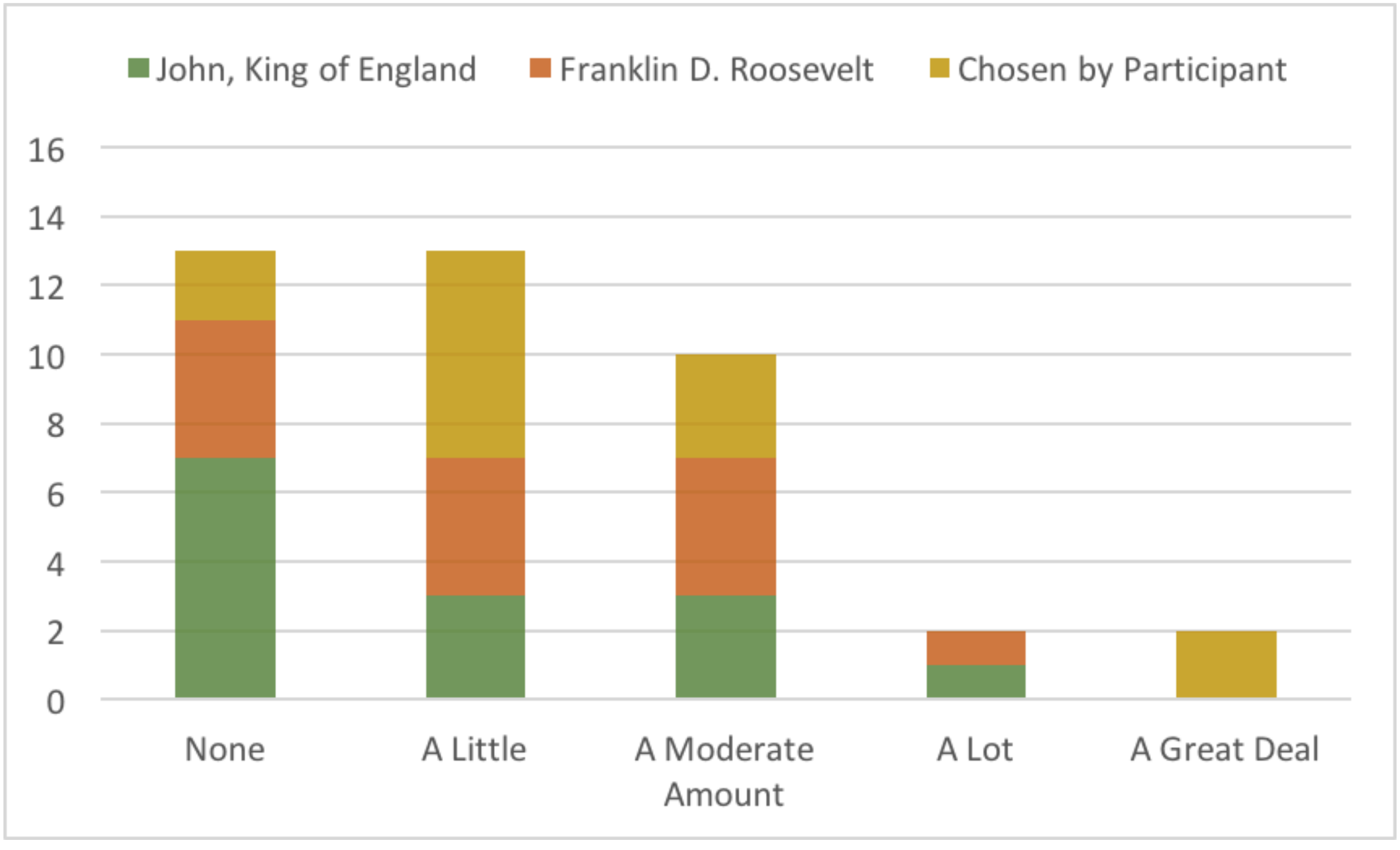
Useful Features



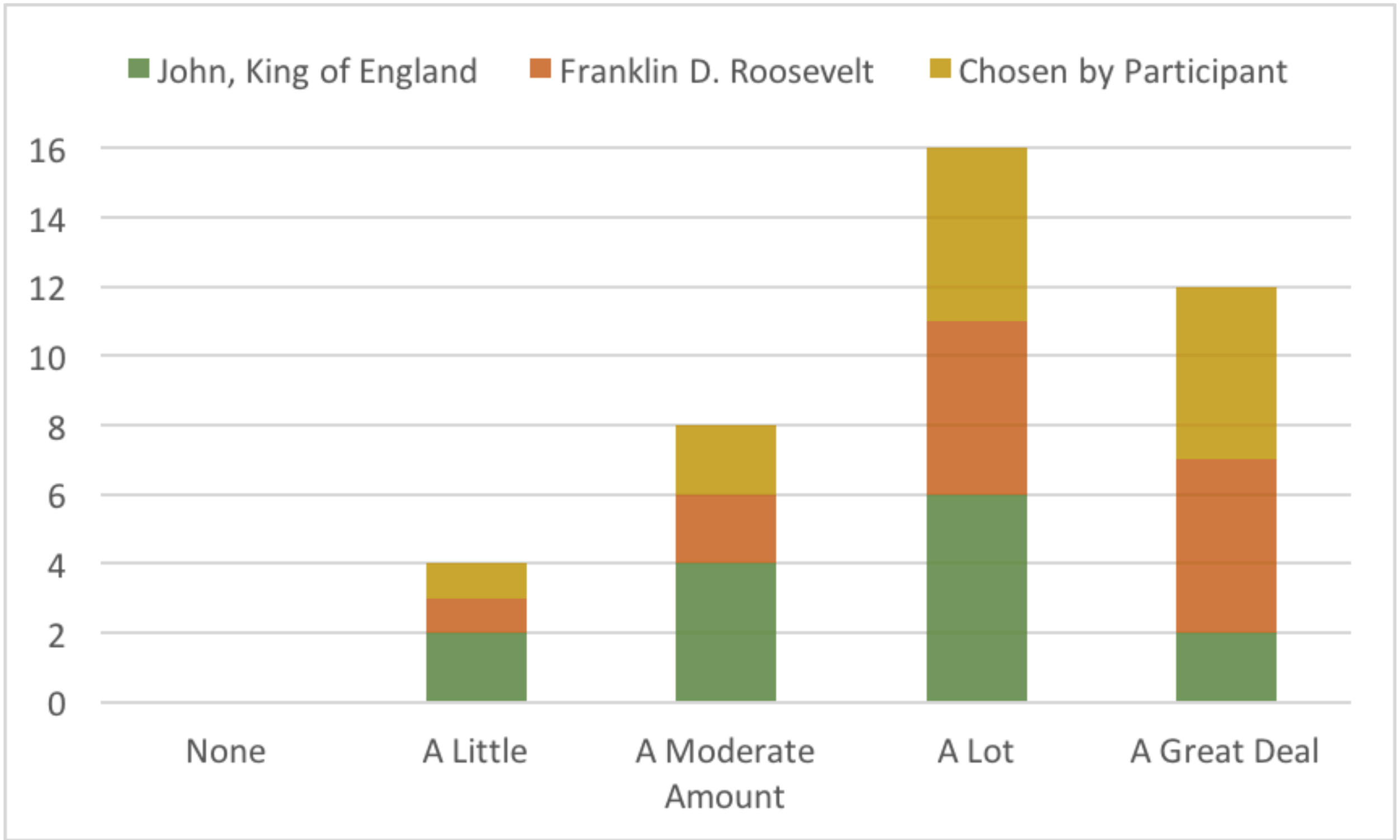
Overall Utility



Prior Knowledge of the User



Knowledge gained by user



Conclusions from user study

- Participants self-reported positive knowledge gain over using text search only.
- Participants found the tool useful and were satisfied
  - Group features that are often used together
  - Re-evaluate the utility of the less-used features
  - Combine storyline, timeline and map for a unified view
  - Produce dynamically generated natural language summaries

Open-source software

github.com/sfu-natlang/lensingwikipedia

$ git clone <repo>
$ docker-compose build
$ docker-compose up
$ make staging
Launch web browser

Thank you!