3/17/2020 Textual Visualization

Why do we need text viz?
- Lots of content outside which are not compiled together. So there is a lot of text out there.
- Efficiently need to acquire and display information. If we can represent the text data vizually it becomes more important.

WordSeer:
- A Text Analysis Environment for Humanities Scholars
- WordSeer is a text analysis environment that combines visualization, information retrieval, sensemaking and natural language processing to make the contents of text navigable, accessible, and useful

Citeology:
- Citeology looks at the relationship between research publications through their use of citations. The names of each of the 3,502 papers published at the CHI and UIST Human Computer Interaction (HCI) conferences between 1982 and 2010 are listed by year and sorted with the most cited papers in the middle. In total, 11,699 citations were made from one article to another within this collection. These citations are represented by the curved lines in the graphic, linking each paper to those that it referenced.

We feel fine project:

Dating websites:
- building reccomendations, matching problem is applicable to many other concepts.
- based on a profile.
- lots of factors to be matched.
- not just about matching the meta data of different people.
- you are hoping to extract the meaning outside the given metadata and then connect users.

foodmood.in:

Pulse of the Nation:
- mood throughout the day, as inferred from twitter.
- sentiment analysis from tweets.

Text viz:
- may converted to numbers
- may converted to viz
- you may spend a lot of time processing the data, decide the viz, but a powerful title, text, summary is needed to append to the vizualization.
- transform text info into visual forms

How to convert text to viz:
- words:
- sentences:
- documents:
- relationships:

What are effective Text Viz:
- content linguistic structure:
- content semantics: sentiment analysis are trying to highlight this
- content similarities: harder to relate, there are no such things that tell you which posts are similar to each other, easiest to compute but harder to project,
- content connections:
- content evolution:
- improve text searches
- innovations:

What to visualize?
- Single document viz:
- Document collection viz:
- Extended document viz:

People are willing to contribute to textual content:
- reviews
- twitter

Machine Learning: Classification techniques:
- Supervised appreach
- Unsupervised approach

Aggregate the network viz to help to explore and identify the relations:

Sentiment viz:
- Sentiment polarity: love <-> hate, thumbs up <-> thumbs down, positive <-> negative. Use sentiment library to annotate and find if text belongs to either of the two extremes?
- Color and vizual elements to represent and highlight the differences.
- Affective words, aspect-specific matter! We can cumulate and transform the text to numbers but this is very naive appraoch (lexicon based). We need to improve potential guarantee. It is challenging. One of the challenges is the domain specific words. eg small etc.
- Indexing words,
- More challenges: noun and verb connotation, multi-word expressions.
- scoring words, phrases, etc with the help of a dictionary.
- Approaches:
    - Lexicon based: word by word assign positive or negative value to all words and form a dictionary.
    - Rule-based:
    - Machine Learning: takes into account any possible features.
    - Hashtags, emoticons
- Challenges from using twitter tweets:
    - Short and incomplete
    - sentences, not documents
    - words can be fuzzy, confusing, abbreviations, informal, compound words etc.
    - cross language mix model
    - implicit sentiment eg: sarcasm, irony, metaphor.

Test Processing:
- Lexicon Analysis:
- Extract Terms: use only some
- Remove Stop words: removing filler words which do not change the meaning of the sentence. like conjunction words, punctuations etc (a, an, the, etc)
- Stemming: standard form of all words to unify all the words and make them singular.
- Term Selection:

Topic Modeling:
- trying to model the topics. Off the shelf libraries, algos which you cna find everywhere.
- documents that have structure.
- LDA:
- Topic Facet Modeling:

Text Viz Lecture 2:

Converte unstructured data to structured data
- Term Document Matrix
- Term -> a single word, could be a phrase "data visualization"
- Document -> A collection of texts
- Each cell (i,j): number of term(i) in document(j)
    - can be binary or count (frequencies)

Engineer Features:
- find out, design the features for out textual content.
-Terms as features
    - Bag of words mechanism: only meaningful words of the content. These set of words are features.
    - n-grams (n consecutive words): contiguous sequence of n items from a given sample of text

- Re-weighting Features:
    - not every word have the same weight, meaning, importance. The adjectives becomes importance in these cases.
    - If a term occurs more frequently in many documnets it has less discriminatory power.
    - Term importance = term frequency (TF) x inverse-documnet frequency (IDF)

Vector Space Model:
- Represent each document by a high dimensional vector in the space of words. Count total number of words in each document.
- Bag of words counts each word's frequency in the document. It does not present the semantics.

Topic Modeling:
- The vector representation does not capture semantic relations between words
- words <->
- The summary shows the semantics. The distribution does this.
- Now we have the numbers of frquency of each word. Next step is viz.

What to Visualize?
- Sinlge Document:
- Collection of Documents
- Extended attribute of Documents

- The vocabulary based viz.
- Semantic Structure:
    - IS-A relation
    - it always carries relations, can represent using netowrk or relations.
    - Subject - verb - object triplets
- Theme based visualization:
    - Trend based collection
- Visualize Document Relations: how you relate to other documents or with words within the same document.
    - Our goal is to build the features
    - Disease, symptoms,treatments are all a bag of words.

Time Series Vizualization

Interactive line chart of live updates
-We can plot various data plots so that we can predict
-If we make it as a predictive models, it can works as well
-We can see mapviews,emojis, etc so we cna aggregate time attributes with other things
-The time spectrum can be used and manipulated to use it for other things

Organize time data with time attribute

- Artifacts in museums are organised by time
- Can incorporate different multivariate data and overlap it.

Series Data:

- Sets of values changing over time
- May have an internal structure : hoping to extract some semantic changes
    - correlation
    - trend
    - seasonal variation
- eg When looking at the computer CPU Utilizations over the period of particular time
    - If you have multiple cpus it becomes very confusing because it becomes very messy. It is very hard to compare point by point. Well you can make it interactive to manage the comparisons easily.

Larger Database

- This shows us that it becomes chalenging if the data set increases highly.
- Can use small multiples technique

Principles:

- Familiar visual representations should be preserved when appropriate
- Side by side comparison of small multiple views is easier that remembering previous seen views
- Overview first,
- SPatial position is the strongest perceptual cue
- Multiple views are most effective when coordinated throught explicit linking
- Avoid abrupt visual changes
- User actions

Why does the discussion matter with time?

- Can uncover a lot of different semantics
- Evidence gathering for detectives and keep a track of the timeline

Topic Evolution Analysis:

- observe the evolution and flow of knowledge in collaboration
- Topic models to summarise a lot of text in semantics

VISA: A visual sentiment analysis system

- Test summarization techniques

Summary:

- visualizing collections of observed events by time
- visualizing observed events that are qualified, classified, examined for possible relations with other facts
    - eg: Behavior Modeling:
        - when: temporal locus
        - how long: temporal extent
        - how many times: repeatability
        - intensity: etc

Behavior changes:

- event series over time
- properties of behavior changes
    - level
    - trend
    - variability
- Time modeling allowed us to observe the behaviour

How to visualiza the Changes over time: by seeing the differences

- magnitude of change:
- shape of change:
- velocity of change: how fast are the changes, use interactions/animations
- direction of change: utilize the colors to highlight it