Cosine Similarity between Questions

1. Import the data into R [every row contains documents]
2. Filter out Stop words
3. Apply Steamer <PORTER, LEMMA>

Introduction to the tm Package Text Mining in R

Visualization

1. Word Cloud: R
2. Directed Graph:
    a. Javascript Library
    Sigma.js
    b. General tool
    Gephi
3. Display: Movie details
4. Utility: Fisheye Distortion

Graph visualization

Sigma.js is compatible with Gephi.

1. import data into Gephi: Gephi has more flexible import data format
    a. Import nodes with size: Data Lab → import spreadsheet
    b. Import edges with weight: File → Open → choose file
        ➔ import the diagonal matrix
        ➔ select “Graph Type” as “Undirected”
        ➔ Gephi is smart enough to only select necessary values
        ➔ No self-loop is allowed in undirected graph
        ➔ Gephi is smart to exclude weight = 0
        ➔ i.e) faraway: gives no meaningful info
2. Brash up the visualization
    a. adjust node size: use ranking in overview
    b. adjust edge length: higher weight closer
        ➔ Gephi uses topology algo Force Atlas QF, QF2
        ➔ Force-based layout algorithms
        ➔ this step is done when importing edges with weight
    c. cannot control edge length and thickness independently QF
    d. Thus use color to represent the link
    e.

3. visualize the data and do some additional algo available in Gephi
    a. cluster by color
    b. K-means: OpenOrd Layout apply all question
    c. etc…
4. Export the result data in GEFX format
5. Import the output into Sigma.js
6. Create HTML container for the graph
7. Host the contents on the web

Memo

- ❏ Excluded include self-loop
- ❏ Don't use *Numbers*: only shows 255 columns
- ❏ Mouse over a node will highlight the other nodes whose weight!=0 i.e) has similarity

Data Acquisition

1. Data Integration: future scope will be combining with other question forum (quora)
2. Data Comprehension
    a. ~~find the document of data source schema~~
    b. ~~provide the overview of each data (available on github)~~
    c. ~~understand the attributes in each data~~
    d. ~~find any noise in data~~
3. ~~Data Cleansing~~ <Command>
   Durable with Perl script but assuming the data is huge use Pig instead
   Tutorial: Hortonworks
   Tutorial: Import XML to Hcatalog, Parsing XML with Pig, Script
   - In posts.xml
    a. remove PostTypeId != 1, 2 --> ~~use Pig~~
    b. extract attribute Body in posts.xml ref
    c. remove symbols for text decoration --> ~~Write perl script~~

Data Analysis

1. Pick an algorithm for analysis
2. ~~Data selection~~
    ✓ find out what fields needed in order to perform that algorithm
        - select certain attributes by using R xml parser
3. ~~Data Integration~~
    ✓ Combine the selected data to generate a new dataset
        - Need to combine every Question with Answers by ParentId
4. ~~Data Transformation~~
    ✓ transform data into forms appropriate for mining if necessary by performing summary or aggregation operations
5. ~~Data Mining~~
    ✓ learn R syntax and semantics to apply the algorithm
6. Pattern evaluation
7. Knowledge presentation
    ✓ Visualize the correlation between questions by network graph
    ✓ Every Question and mouse hover to display the Accepted answer
    ✓