**CSDS 234 Structured and Unstructured Data  
Fall 2024**

Assignment 5



1. [**Information extraction**] **(20).** This question is a practice of a “manual” conversion from a given text to an RDF graph. Below is a paragraph about Tim Berners-Lee (https://dbpedia.org/page/Tim\_Berners-Lee).

Write a set of RDF triples in Turtle format to encode the facts about him. It is not required that you encode all the details, but major facts that can be recognized as “RDF triples”.

“Sir Timothy John Berners-Lee OM KBE FRS FREng FRSA FBCS (born 8 June 1955), also known as TimBL, is an English computer scientist best known as the inventor of the World Wide Web. He is a Professorial Fellow of Computer Science at the University of Oxford and a professor at the Massachusetts Institute of Technology (MIT). Berners-Lee proposed an information management system on 12 March 1989, then implemented the first successful communication between a Hypertext Transfer Protocol (HTTP) client and server via the Internet in mid-November. He is a senior researcher and holder of the 3Com founder's chair at the MIT Computer Science and Artificial Intelligence Laboratory (CSAIL). He received the 2016 Turing Award. ”

[*If you want to find some clue, try “Browse using” on the same page*] ☺

A set of RDF triples in turtle:

@prefix dbr: <http://dbpedia.org/resource/> .

@prefix dbo: <http://dbpedia.org/ontology/> .

@prefix dbp: <http://dbpedia.org/property/> .

@prefix foaf: <http://xmlns.com/foaf/0.1/> .

dbr:Tim\_Berners-Lee

a dbo:Scientist ;

foaf:name "Timothy John Berners-Lee" ;

foaf:nickname "TimBL" ;

dbo:birthDate "1955-06-08"^^xsd:date ;

dbo:birthPlace dbr:England ;

dbo:nationality dbr:England ;

dbo:knownFor dbr:World\_Wide\_Web ;

dbo:occupation "Computer Scientist" ;

dbo:award dbr:Turing\_Award ;

dbo:awardYear "2016"^^xsd:gYear ;

dbo:position [

a dbo:Position ;

dbp:title "Professorial Fellow of Computer Science at the University of Oxford"

] ;

dbo:position [

a dbo:Position ;

dbp:title "Professor at the Massachusetts Institute of Technology"

] ;

dbo:position [

a dbo:Position ;

dbp:title "Senior Researcher and holder of the 3Com founder's chair at MIT CSAIL"

] ;

dbo:proposal [

a dbo:Proposal ;

dbo:proposedOn "1989-03-12"^^xsd:date ;

dbo:proposedFor dbr:Information\_management\_system

] ;

dbo:accomplishment [

a dbo:Accomplishment ;

dbp:event "First successful communication between an HTTP client and server" ;

dbp:date "1989-11"^^xsd:gYearMonth

] .

1. **[MapReduce] (30) Design MR algorithms [describe the Map and Reduce functions, and the MR algorithm].** 
   1. *Top‐k keywords.* Search engines like Google maintains popular webpages in a set R for keyword search. Each record r in R is an article, stored as a sequence of keywords. Given the set R, describe a MapReduce algorithm to report the top k most frequent keywords appeared in the webpages in R. Describe your MR program. (hint: solve a word count problem).

Task: Identify the top k most frequent words

MR algorithm:

* Map Function:
  + Input: A record r from R (an article stored as a sequence of keywords)
  + Output: A key-value pair of each keyword with the value as 1
    - Ex: (keyword, 1)
* Reduce Function:
  + Input: A keyword and the counts from all of the mappers
    - Ex: (keyword, [1,1,1,1,1]
  + Output: A single keyword and the total count
    - Ex: (keyword, 5) where 5 is the total count
* Post-Processing:
  + The reducer outputs a list of all keywords and total count pairs. After that, you want to sort the list in descending order and take the top k.

Describe the MR Program:

* 1. *Common friends.*A social network provider updates the “common friends” between two users and response to hundreds of millions of such queries every day. The friendship information is stored as key-value pairs (Person, [List of friends]) for every user. Describe a MapReduce program to return a dictionary of common friends of the form ((User i, User j), [List of Common friends of i and j]) for all pairs of i and j who are friends. The order of i and j you returned should be the same as the lexicographical order of their names. Describe the Map() and Reduce() function.

Task: Determine the common friends of all pairs (Friendi, Friendj) where i > j

MR algorithm:

* Map Function:
  + Input: A key-value pair
  + Output: Pairs of friends in alphabetical order with the mutual friend
* Reduce Function:
  + Input: A pair of (useri, userj) and a list of their friends [List1, List2]
  + Output: The pair of users and a list of their intersection of common friends

1. **[MapReduce] (30) Design and describe MR algorithms [No need to give detailed code; just provide Map and Reduce function following the examples in the lecture notes].** 
   1. *Vote-count*. In an election activity, a large number of votes need to be counted. A vote is a key-value pair (vote id, candidate id). For example, (2, 1) means the second voter/ballot voted for candidate 1. Describe an MR algorithm that counts how many unique votes of each candidate there are (for example, it produces key-value pairs (candidate id, count) for each candidate).

* Map Function:
  + Input: (vote id, candidate id)
    - Each input record represents a vote with the voter ID and candidate ID
  + Output: (candidate id, 1)
  + Process: Emit the candidate id as the key and 1 as the value for every vote
* Reduce Function:
  + Input: (candidate id, [1,1,1,…]
    - Aggregated values of each of the candidates from the map function
  + Output: (candidate id, count)
  + Process: Sum of all of the values of each of the candidates by ID for the total votes
  1. *Top‐k frequent words.* Search engines often maintain popular web pages and retrieve most frequent keywords to support fast keyword search. Consider a set R of web pages. Each record r from set R is in the form of <docid, [List of terms]>, which contains the id (docid) of a web page, and a list of terms in the web page (split by spaces). Below are two examples:   
     <1, [Harry, Potter, fantasy, novels, lives, wizard]> and <2, [Harry, young, wizard, Harry, student, Hogwarts]>.
* Map Function:
  + Input: (docid, [List of terms])
    - Each input record is a document ID and its list of terms
  + Output: (term, 1)
  + Process: For each of the items in the list, emit the term as the key and the value as 1
* Reduce Function:
  + Input: (term, [1,1,1,1,…]
    - Aggregated values for each term from the Map function
  + Output: (term, count)
  + Process: Sum all of the values for each of the terms to get the frequency of the term

Given this set R, describe a MapReduce algorithm to report the top k most frequent terms and their frequency as appeared in the web pages in R. For the above example, the output returns top 2 most frequent terms as <Harry, 3> and <wizard, 2>.

* Post-processing:
  + Input: (term, count)
  + Process: Extract the top k terms after it is sorted in descending order

1. **Stream Data (10).**
   1. Describe how streaming data needs to be handled differently in a database and describe a “case study” example of streaming data.

* Streaming data is the continuous flow of data that arrives in real time (or close enough to real time) from a source, whether that is a sensor, social media feeds, or streams. To handle streaming data, you need to process as you go. This is different from traditional databases, which can process in batches of fixed sizes. In streaming databases, you might not need all the data. Maybe you need a certain time that something happens. Traditional databases store data in tables that are fixed, while streaming databases may not have a way to store data in a fixed way.
* A scenario of streaming data:
  + Social media data: A social media company has access to the streaming data of their brand from multiple social media sites. They can see when people have tagged them and when they have posted, comments on posts, likes, etc.
  + All of the information is public and can be pulled from an API or through scraping
  + If there are negative words coming up, the brand can act on the backlash that they will have.
  + Let’s say that a brand releases a new pair of sweatpants. They post on their story with a hashtag the reference the sweatpants. They say in the post that they encourage people to use the hashtag. They also make sure that the picture of the sweatpants is unique. They first post on Instagram, and the system processes the comments. There are a mix of negative and positive comments. The system can process which of the comments are negative and which are positive as well as how negative or positive they are. There is a score attached to the post, overall positive or negative. Using this analysis, the brand can see how the product has been perceived by people, whether negatively or positively.

1. **Data Visualization.**
   1. Describe a data set of your choosing. Describe the process you would choose to visualize that data for

* Dataset: Customer Feedback Analysis. The dataset contains customer feedback data collected from an ecommerce website and uses their reviewing system. Each of the record has the following information: Customer ID, Product ID, Review Text, Rating (scale of 1-5), Date Review was submitted, Sentiment analysis of review (-1 to 1 scale), Region of customer
  + 1. yourself to clean the data
  + Histogram of Rating and Sentiment Analysis
  + A box plot that detects outliers in rating and sentiment analysis
  + Heatmap of correlated figures
    1. for your manager to convince them that you have the correct answer
  + Line chart that shows how products are trending and how the company overall is trending
  + Scatterplot between rating and sentiment analysis
    1. for the CEO of the company that your team is hitting deliverables and moving in the correct direction
  + Line chart on trend in rating by year and/or month
  + Bar chart for regions with positive or negative reviews (places to target)