# CSIT 6000I Search Engines and Applications

# Mid-Term Examination II, Fall 2019

# 23 Nov, 2019, 1:30pm

# Time Allowed: 24 hours

**Rule:** You can use any tool or information source to answer the questions. You are allowed to be online and use any tools or information sources to help you with answering the questions. The only restriction is that YOU MUST DO IT BY YOURSELF AND NOT CONSULT OR DISCUSS WITH ANY PEOPLE.

1.[40] You have the following documents, and important keywords to be indexed are underlined.

|  |  |
| --- | --- |
| D1 | I use Windows and Azure Cloud from Microsoft. |
| D2 | Microsoft Chairman Bill Gates announced retirement this morning. |
| D3 | Success of Windows is attributed to Gates. |
| D4 | Azure Cloud becomes successful after Gates left Microsoft. |
| D5 | I watched sunrise in my garden. There was a little cloud in the morning. |
| D6 | I like windows with wooden frames. |
| D7 | I have wooden windows and wooden gates. |
| D8 | Through the windows, I can see trees and flowers in my garden. |

Using R, Python or any programming language or online websites you are familiar with, answer the following questions. You can include the screenshot, or copy-and-paste the output into your answer.

1. [5] Create the term-document matrix for the document set; use tf as the term weight, and normalize the document vector lengths to 1.
2. [5] Perform SVD on the matrix and show the USV matrices.
3. [5] Perform Rank-2 approximation and plot the documents on a two-dimensional space.
4. [10] Map the two queries (i) Microsoft windows, and (ii) windows garden, into the reduced vector space and plot them on the same two-dimensional graph as in (c).
5. [5] Using inner product, compute the top 3 documents for each query in the reduced vector space; compare them with the top 3 documents in the original vector space, i.e., directly from the term-document matrix in (a).
6. [10] Discuss how synonymy and polysemy are resolved, or are not resolved, with LSI in this example.

2. [20] Using the document set in Q. 1, identify the two words that you would choose as index terms using term discrimination value.

3. [20] A search for trade conflict returns the following results. The check marks mean the user clicked the corresponding pages; unchecked pages are not clicked. The user clicked Page 2 before Page 8. Show the page preference pairs derived from each of the five preference mining strategies discussed in the lecture.

|  |  |  |
| --- | --- | --- |
|  | China–United States trade war - Wikipedia |  |
|  | Trade war - Wikipedia | √ |
|  | The Impact of Trade Conflict on Developing Asia | Asian ... |  |
|  | US-China trade dispute | Financial Times |  |
|  | US-China trade war | South China Morning Post |  |
|  | The US-China Trade War: A Timeline - China Briefing News |  |
|  | A quick guide to the US-China trade war - BBC News |  |
|  | Trading Away from Conflict: Using Trade to Increase ... | √ |
|  | China and the United States: Trade Conflict and Systemic ... |  |

4.[10] The web consists of a set of linked web pages; search engine is a software which directs users to a small subset of the webpages (typically 10 pages for each query). Discuss how the existence of Google, which implements PageRank and indexes all pages on the web, impacts the random surfer model, in which users first pick a random page to visit and after landing on the page, either randomly pick another page to visit or following a random link in that page. Write no more than 5 sentence.