CODE CHALLENGE WIKI FOR WWI

Part 1

For the Wikipedia page on World War 1,

https://en.wikipedia.org/wiki/World\_War\_I, download and store the main

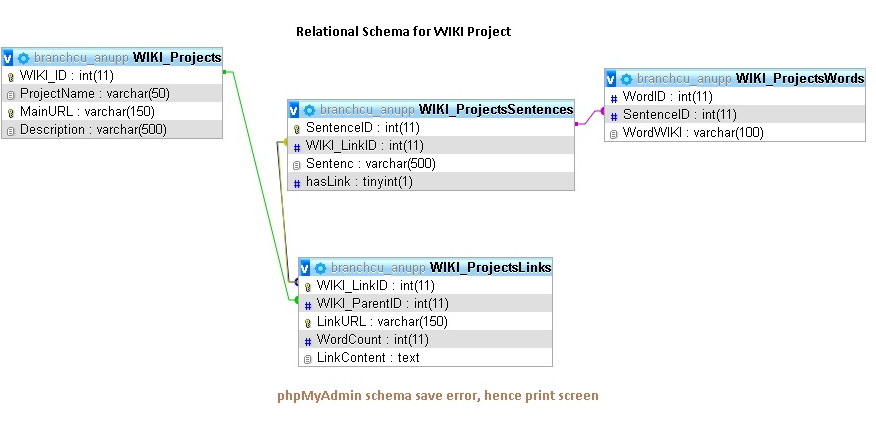
and all linked pages (just go one level in). From this dataset, design

and create a data model for answering the questions below. Load the

data into the data model to answer the questions below using the SQL

flavor of your choice.

**Response:**



Part 2

**Q1) What are the top 10 most used words on these pages and the total**

**number of times they appear? Exclude articles and conjunctions.**

**Response**

1) War: 4341 times

2) German: 4091

3) Army: 3028

4) British: 3017

5) French: 2180

6) World: 1879

7) City: 1877

8) Ottoman: 1853

9) Battle: 1665

10) Division: 1631

**Q2) What are the top 10 countries referenced on these pages and what**

**is the average length of the sentence they appear in?**

**Response**

| [**count(WordWIKI)**](https://securem72.sgcpanel.com:2083/cpsess1633335780/3rdparty/phpMyAdmin/sql.php?db=branchcu_anupp&table=ltb_countries&sql_query=Select+count%28WordWIKI%29%2C+Avg%28Char_length%28Sentenc%29%29%2C+Country+from+ltb_countries+ltc+inner+join+WIKI_ProjectsWords+WW+on+ltc.Country+%3D+WW.WordWiki+inner+join+WIKI_ProjectsSentences+Sen+on+WW.SentenceID%3DSen.SentenceID+group+by+Country+%0AORDER+BY+%60count%28WordWIKI%29%60+ASC+limit+30&session_max_rows=25&is_browse_distinct=0) | [**Avg(Char\_length(Sentenc))**](https://securem72.sgcpanel.com:2083/cpsess1633335780/3rdparty/phpMyAdmin/sql.php?db=branchcu_anupp&table=ltb_countries&sql_query=Select+count%28WordWIKI%29%2C+Avg%28Char_length%28Sentenc%29%29%2C+Country+from+ltb_countries+ltc+inner+join+WIKI_ProjectsWords+WW+on+ltc.Country+%3D+WW.WordWiki+inner+join+WIKI_ProjectsSentences+Sen+on+WW.SentenceID%3DSen.SentenceID+group+by+Country+%0AORDER+BY+%60Avg%28Char_length%28Sentenc%29%29%60+ASC+limit+30&session_max_rows=25&is_browse_distinct=0) | [**Country**](https://securem72.sgcpanel.com:2083/cpsess1633335780/3rdparty/phpMyAdmin/sql.php?db=branchcu_anupp&table=ltb_countries&sql_query=Select+count%28WordWIKI%29%2C+Avg%28Char_length%28Sentenc%29%29%2C+Country+from+ltb_countries+ltc+inner+join+WIKI_ProjectsWords+WW+on+ltc.Country+%3D+WW.WordWiki+inner+join+WIKI_ProjectsSentences+Sen+on+WW.SentenceID%3DSen.SentenceID+group+by+Country+%0AORDER+BY+%60ltc%60.%60Country%60+ASC+limit+30&session_max_rows=25&is_browse_distinct=0) |  |
| --- | --- | --- | --- |
| 1236 | 179.4676 | Germany |  |
| 901 | 177.7725 | France |  |
| 789 | 170.3308 | Russia |  |
| 538 | 173.8755 | Italy |  |
| 420 | 178.8381 | Britain |  |
| 348 | 175.9023 | Serbia |  |
| 337 | 180.3858 | Bosnia |  |
| 313 | 209.8147 | Jordan |  |
| 298 | 162.8389 | Ireland |  |
| 296 | 183.8986 | Austria |  |

**SQL USED:**

Select count(WordWIKI), Avg(Char\_length(Sentenc)), Country from ltb\_countries ltc inner join WIKI\_ProjectsWords WW on ltc.Country = WW.WordWiki inner join WIKI\_ProjectsSentences Sen on WW.SentenceID=Sen.SentenceID group by Country order by count(wordWIKI) desc limit 30

**Q3) For the main WW1 page, estimate how many words are used to**

**describe each year of the war (1914 – 1918) and the top 5 locations**

**referenced for each year? \*Bonus, test the code on the main WW2 page**

**and see if the results make sense given the differences and years in**

**the wars.**

**SQL Used**

| [**Sum(Char\_Length(Sentenc))**](https://securem72.sgcpanel.com:2083/cpsess1633335780/3rdparty/phpMyAdmin/sql.php?db=branchcu_anupp&table=ltb_countries&sql_query=Select+Sum%28Char_Length%28Sentenc%29%29%2C+WORDWIKI+from+%28Select+SentenceID%2C+WordWiki+from+WIKI_ProjectsWords+where+wordWIKI+IN+%28%271914%27%2C+%271915%27%2C+%271916%27%2C+%271917%27%2C+%271918%27%29%29+WW+INNER+join+WIKI_ProjectsSentences+Sen+on+WW.SentenceID+%3D+Sen.SentenceID+where+WIKI_Linkid+%3D1+group+by+WORDWIKI++%0AORDER+BY+%60Sum%28Char_Length%28Sentenc%29%29%60+ASC&session_max_rows=25&is_browse_distinct=0) | [**WORDWIKI**](https://securem72.sgcpanel.com:2083/cpsess1633335780/3rdparty/phpMyAdmin/sql.php?db=branchcu_anupp&table=ltb_countries&sql_query=Select+Sum%28Char_Length%28Sentenc%29%29%2C+WORDWIKI+from+%28Select+SentenceID%2C+WordWiki+from+WIKI_ProjectsWords+where+wordWIKI+IN+%28%271914%27%2C+%271915%27%2C+%271916%27%2C+%271917%27%2C+%271918%27%29%29+WW+INNER+join+WIKI_ProjectsSentences+Sen+on+WW.SentenceID+%3D+Sen.SentenceID+where+WIKI_Linkid+%3D1+group+by+WORDWIKI++%0AORDER+BY+%60WORDWIKI%60+ASC&session_max_rows=25&is_browse_distinct=0) |  |
| --- | --- | --- |
| 3020 | 1914 |  |
| 3736 | 1915 |  |
| 3410 | 1916 |  |
| 6056 | 1917 |  |
| 6397 | 1918 |  |

[Select](https://securem72.sgcpanel.com:2083/cpsess1633335780/3rdparty/phpMyAdmin/url.php?url=https://dev.mysql.com/doc/refman/5.5/en/select.html) [Sum](https://securem72.sgcpanel.com:2083/cpsess1633335780/3rdparty/phpMyAdmin/url.php?url=https://dev.mysql.com/doc/refman/5.5/en/group-by-functions.html#function_sum)(Char\_Length(Sentenc)), WORDWIKI from ([Select](https://securem72.sgcpanel.com:2083/cpsess1633335780/3rdparty/phpMyAdmin/url.php?url=https://dev.mysql.com/doc/refman/5.5/en/select.html) SentenceID, WordWiki from WIKI\_ProjectsWords where wordWIKI [IN](https://securem72.sgcpanel.com:2083/cpsess1633335780/3rdparty/phpMyAdmin/url.php?url=https://dev.mysql.com/doc/refman/5.5/en/comparison-operators.html#function_in) ('1914', '1915', '1916', '1917', '1918')) WWINNER join WIKI\_ProjectsSentences Sen on WW.SentenceID = Sen.SentenceID inner join WIKI\_ProjectsLinks WL on WL.WIKI\_LinkID= Sen.WIKI\_LinkID whereLinkURL='/wiki/World\_War\_I' group by WORDWIKI

**Top 5 locations for each year:**

| [**contcount**](https://securem72.sgcpanel.com:2083/cpsess1633335780/3rdparty/phpMyAdmin/tbl_sql.php?db=branchcu_anupp&table=WAgain#usesubform[2]=1) | [**CNTRY**](https://securem72.sgcpanel.com:2083/cpsess1633335780/3rdparty/phpMyAdmin/tbl_sql.php?db=branchcu_anupp&table=WAgain#usesubform[3]=1) | [**WWI\_Years**](https://securem72.sgcpanel.com:2083/cpsess1633335780/3rdparty/phpMyAdmin/tbl_sql.php?db=branchcu_anupp&table=WAgain#usesubform[4]=1) |  |
| --- | --- | --- | --- |
| 2 | Britain | 1914 |  |
| 2 | Ireland | 1914 |  |
| 1 | Canada | 1914 |  |
| 1 | Serbia | 1914 |  |
| 1 | America | 1914 |  |
| 4 | Italy | 1915 |  |
| 3 | Germany | 1915 |  |
| 3 | Britain | 1915 |  |
| 2 | France | 1915 |  |
| 1 | England | 1915 |  |
| 1 | Germany | 1916 |  |
| 1 | Macedonia | 1916 |  |
| 1 | Britain | 1916 |  |
| 1 | Ireland | 1916 |  |
| 1 | England | 1916 |  |
| 3 | Russia | 1917 |  |
| 2 | France | 1917 |  |
| 1 | Latvia | 1917 |  |
| 1 | Germany | 1917 |  |
| 1 | Estonia | 1917 |  |
| 24 | Germany | 1918 |  |

(Select count(WAgain.WordWIKI) as contcount, WAgain.WordWIKI as CNTRY, WWI\_Years from (Select SentenceID, WordWiki as WWI\_Years from WIKI\_ProjectsWords where wordWIKI = ('1914')) WW INNER join WIKI\_ProjectsSentences Sen on WW.SentenceID = Sen.SentenceID inner join WIKI\_ProjectsLinks WL on WL.WIKI\_LinkID= Sen.WIKI\_LinkID inner join WIKI\_ProjectsWords WAgain on Sen.SentenceID = WAgain.SentenceID inner join ltb\_countries ltbc on ltbc.Country=WAgain.WordWIKI

where LinkURL='/wiki/World\_War\_I' group by WAgain.WordWIKI order by count(WAgain.WordWIKI) desc limit 5)

Union All

(Select count(WAgain.WordWIKI) as contcount, WAgain.WordWIKI as CNTRY, WWI\_Years from (Select SentenceID, WordWiki as WWI\_Years from WIKI\_ProjectsWords where wordWIKI = ('1915')) WW INNER join WIKI\_ProjectsSentences Sen on WW.SentenceID = Sen.SentenceID inner join WIKI\_ProjectsLinks WL on WL.WIKI\_LinkID= Sen.WIKI\_LinkID inner join WIKI\_ProjectsWords WAgain on Sen.SentenceID = WAgain.SentenceID inner join ltb\_countries ltbc on ltbc.Country=WAgain.WordWIKI

where LinkURL='/wiki/World\_War\_I' group by WAgain.WordWIKI order by count(WAgain.WordWIKI) desc limit 5)

Union all

(Select count(WAgain.WordWIKI) as contcount, WAgain.WordWIKI as CNTRY, WWI\_Years from (Select SentenceID, WordWiki as WWI\_Years from WIKI\_ProjectsWords where wordWIKI = ('1916')) WW INNER join WIKI\_ProjectsSentences Sen on WW.SentenceID = Sen.SentenceID inner join WIKI\_ProjectsLinks WL on WL.WIKI\_LinkID= Sen.WIKI\_LinkID inner join WIKI\_ProjectsWords WAgain on Sen.SentenceID = WAgain.SentenceID inner join ltb\_countries ltbc on ltbc.Country=WAgain.WordWIKI

where LinkURL='/wiki/World\_War\_I' group by WAgain.WordWIKI order by count(WAgain.WordWIKI) desc limit 5)

Union all

(Select count(WAgain.WordWIKI) as contcount, WAgain.WordWIKI as CNTRY, WWI\_Years from (Select SentenceID, WordWiki as WWI\_Years from WIKI\_ProjectsWords where wordWIKI = ('1917')) WW INNER join WIKI\_ProjectsSentences Sen on WW.SentenceID = Sen.SentenceID inner join WIKI\_ProjectsLinks WL on WL.WIKI\_LinkID= Sen.WIKI\_LinkID inner join WIKI\_ProjectsWords WAgain on Sen.SentenceID = WAgain.SentenceID inner join ltb\_countries ltbc on ltbc.Country=WAgain.WordWIKI

where LinkURL='/wiki/World\_War\_I' group by WAgain.WordWIKI order by count(WAgain.WordWIKI) desc limit 5)

Union all

(Select count(WAgain.WordWIKI) as contcount, WAgain.WordWIKI as CNTRY, WWI\_Years from (Select SentenceID, WordWiki as WWI\_Years from WIKI\_ProjectsWords where wordWIKI = ('1918')) WW INNER join WIKI\_ProjectsSentences Sen on WW.SentenceID = Sen.SentenceID inner join WIKI\_ProjectsLinks WL on WL.WIKI\_LinkID= Sen.WIKI\_LinkID inner join WIKI\_ProjectsWords WAgain on Sen.SentenceID = WAgain.SentenceID inner join ltb\_countries ltbc on ltbc.Country=WAgain.WordWIKI

where LinkURL='/wiki/World\_War\_I' group by WWI\_Years order by count(WAgain.WordWIKI) desc limit 5)

**Part 3**

Q1) Using a scripting language of your choice and the raw html page

data, provide a list of all pages where the country “Turkey” is

referenced and estimate the number of words on the page.