QTM 446W

HW5

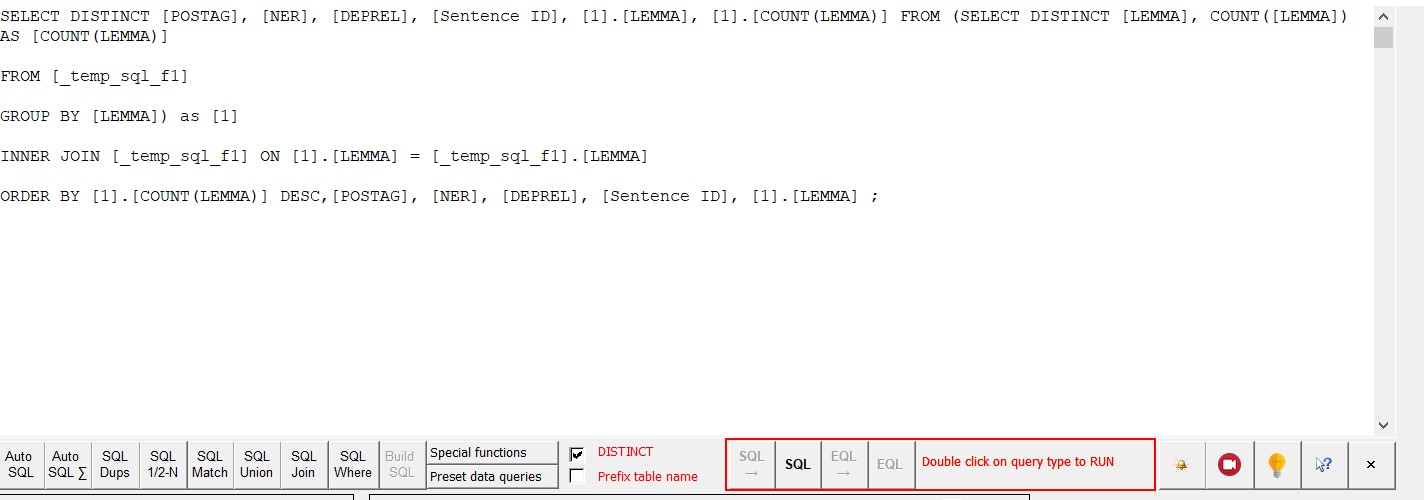
Gabriel Wang

An N-gram is a sequence of n 1-gram. It displays the pattern of a sequence of words. This can reveal very useful information beyond micro level if we consider the higher-level properties such as news tone and geographic location (Leetaru,2011). This is recently made available via various scanning technology and Google digitalized library (Michel et al,2011). However, the accuracy of this kind of important metadata such as dates and categories under suspicion, and someone may call this quantitative analysis method a ‘disaster’ rather a shortcut (Nunberg,2009). The worries are not groundless. The work of Michel and his colleagues shows that the use of ‘briet’ peaked in the 1950s which was a misspelling of ‘brief’. However, this was caused by the scanning error of misrecognizing italic ‘f’ to ‘t’. Thus, we will need reliable metadata and good frequent words.

To address meaningful questions about my corpus, I first decided to determine which words are the significant words for the book *A Song of Ice and Fire* (addressed as SIF). The first step was to clean up the CoNLL table to exclude any punctuations, auxiliary words, special characters, modal verbs and all numeric values via PC-ACE filter tool and aggregate the count of the most frequent words to generate potential 1-gram significant words. To visualize the 1-gram, I wanted to know the sentence id cutoff for each book, and this was made available by inserting the author’s name at the beginning of each book prior the NLP. The sentence id cutoff is recorded as:

|  |  |
| --- | --- |
| Book | Sentence ID |
| A Game of Throne | 4-37258 |
| A Clash of Kings | 37258-78796 |
| A Storm of Swords | 78796-113459 |

The query:

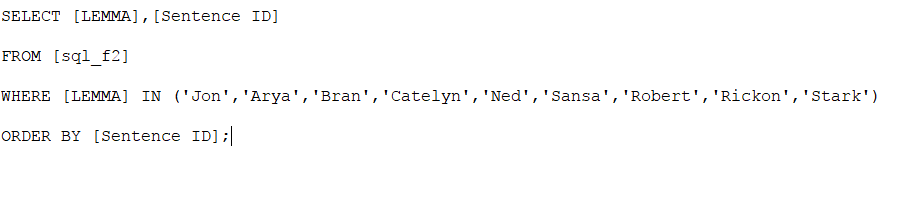


The result for the top 20 meaningful significant words were identified as:

Lord: 3572; Ser: 2454; Jon 1462; Tyrion: 1458; Father: 1384; Brother: 1248; Arya: 1054; Bran: 1033; sword: 949; horse: 945; Wall: 905; Catelyn: 887; Ned: 872; Sansa: 844; Robb: 786; Stark: 750; Maester: 747; Knight: 744; Watch: 724; God: 705;

The most frequent words were character names form those big houses, most of them from House Stark. After I considered the remaining top significant words, I realized that they are either names, places, titles or organizations. So, I figured it probably is worthy a study to focus on comparing the occurrence of significant words of the same type and visualize them. In this report, I used the search and visualization tools to display the occurrences of three great families: House Stark, House Lannister, House Targaryen. I obtained the main members’ first names of these houses with some web scrapping work and integrated the count of these first names occurrences with that of their House name. The result is as follow:

It is easy to tell that as the principle family with the most POV characters, House Stark, dominates the other two Houses not only in number of total appearances but also number of main characters. This is a partial result as some of the characters were always addressed with a nick name, for example, ‘the Imp’ for Tyrion Lannister or ‘Cat’ for Catelyn Stark. Indeed this is only a very small set of characters appeared in SIF, but this could be a good start to the bigger project to visualizing the whole story to the map of Game of Thrones.

Furthermore, I explored the occurrences of these great houses in different books again with the SQL query. 

This time I integrated the whole house instead of separate characters. The visualized result is displayed as follow:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Stark | Lannister | Targaryen |
| Book\_1 | 4396 | 1421 | 413 |
| Book\_2 | 2556 | 1430 | 176 |
| Book\_3 | 1737 | 1482 | 148 |

This chart result is probably shocking to someone who doesn’t read or watch Game of Throne. The author George R. R. Martin is very famous for ‘killing’ the leading characters in his books. And in fact, the reason why the occurrences of House Stark experienced a drop in Book 2 from Book 1 is that Martin, beheaded Lord Ned Stark at the end of book one and Ned himself appeared in over 2000 sentences. House Targaryen’s story didn’t take place in the mainland Westeros so Dany’s name rarely crossed with other characters’ story line which explained why the occurrences are low and steady.

The google N-gram viewer used a broader corpus than the user-defined PC-ACE N-gram viewer. Apparently, google N-gram viewer didn’t include Song of Ice and Fire as it didn’t have any data for ‘Lannister’ and ‘Targaryen’. The PC-ACE runs the n-gram search based on corpus we provided. A similar routine of n-grams always indicates that the two corpus has something similar.