**Functional requirements**

The system should suggest top N(let’s say top ten) frequent and relevant terms to the user based on the text a user types in the search box.

**Non-functional requirements**

* Low latency: The system should show all the suggested queries in real time after a user types. The latency shouldn’t exceed 200 ms. A study suggests that the average time between two keystrokes is 160 milliseconds. So, our time-budget of suggestions should be greater than 160 ms to give a real-time response. This is because if a user is typing fast, they already know what to search and might not need suggestions. At the same time, our system response should be greater than 160 ms. However, it should not be too high because in that case, a suggestion might be stale and less useful.
* Fault tolerance: The system should be reliable enough to provide suggestions despite the failure of one or more of its components.
* Scalability: The system should support the ever-increasing number of users over time.