Data Collection:

Our data collection strategy was designed to capture a broad spectrum of perspectives on our chosen topics, leveraging three major Chinese social media platforms: Weibo, Weixin, and Zhihu. For Weibo, we utilized a dedicated open-source spider to systematically extract posts containing our specified keywords. In the case of Weixin, we employed Python's Selenium module to automate keyword searches within the Sogou browser, specifically targeting Weixin official accounts. This approach facilitated the comprehensive collection of relevant articles, including their titles, authors, and full text. For Zhihu, we adopted a similar strategy to Weixin, automating keyword searches and applying filters based on the number of thumbs up and publish time. This robust data collection methodology enabled us to compile a rich dataset comprising articles, answers, and comments.

Text Analysis:

Upon completion of data collection, we undertook a detailed text analysis using Python's Natural Language Processing (NLP) modules. The first step involved cleaning the text data, which included the removal of punctuation, conversion of all text to lower case, and elimination of common Chinese 'stop words' such as “的”. We then tokenized the cleaned text into individual words or 'tokens', facilitating the analysis of word frequency across our entire dataset. To compile a frequency distribution of the words, we employed Python's Counter module, an efficient tool for tallying the occurrence of objects such as words. The most frequently occurring words were subsequently visualized using a word cloud, offering a clear and impactful representation of the prevalent themes within our dataset. This process not only provided valuable insights into the dominant themes but also informed our subsequent, more detailed analyses.