Before you begin, you may notice a file called generateData.py. This file is used to create a whole folder of data consisting of individual articles from the query specified in the file. This folder, along with maxNumberOfFiles and trigger processing time allows for emulation of Real-Time Streaming because our free plan for NewsAPI does not allow for RTS and gathering of more than 100 articles at a time. You do not need to run generateData.py! The data is present in articles/ for your viewing and you would need to generate an API key to run the generation.

Additionally, the search query we use is "apple".

This search query and the command in generateData.py will obtain all articles allowed from the oldest time our plan allows to the newest time our plan allows. Information obtained includes “author”, “title”, “description”, “content”, and “publishedAt”. The Named Entity Recognition is dependent on the information parsed from the first 4 fields mentioned above (everything but publishedAt).

Results:

The kibana bar chart results indicate the Named Entity counts after 5, 15, and 25 minutes of time. As you can see, the words shown are clearly relevant to Apple, apple being the first most prominent entry. It also has words associated with their products like “iphone”, “repair”, “pro”, and “tv” and words associated with sales features like “deal”, “iphone”, “friday”, “15”, and “new”. These words are all relevant to our search query and appear in increasingly populous values as time progresses.

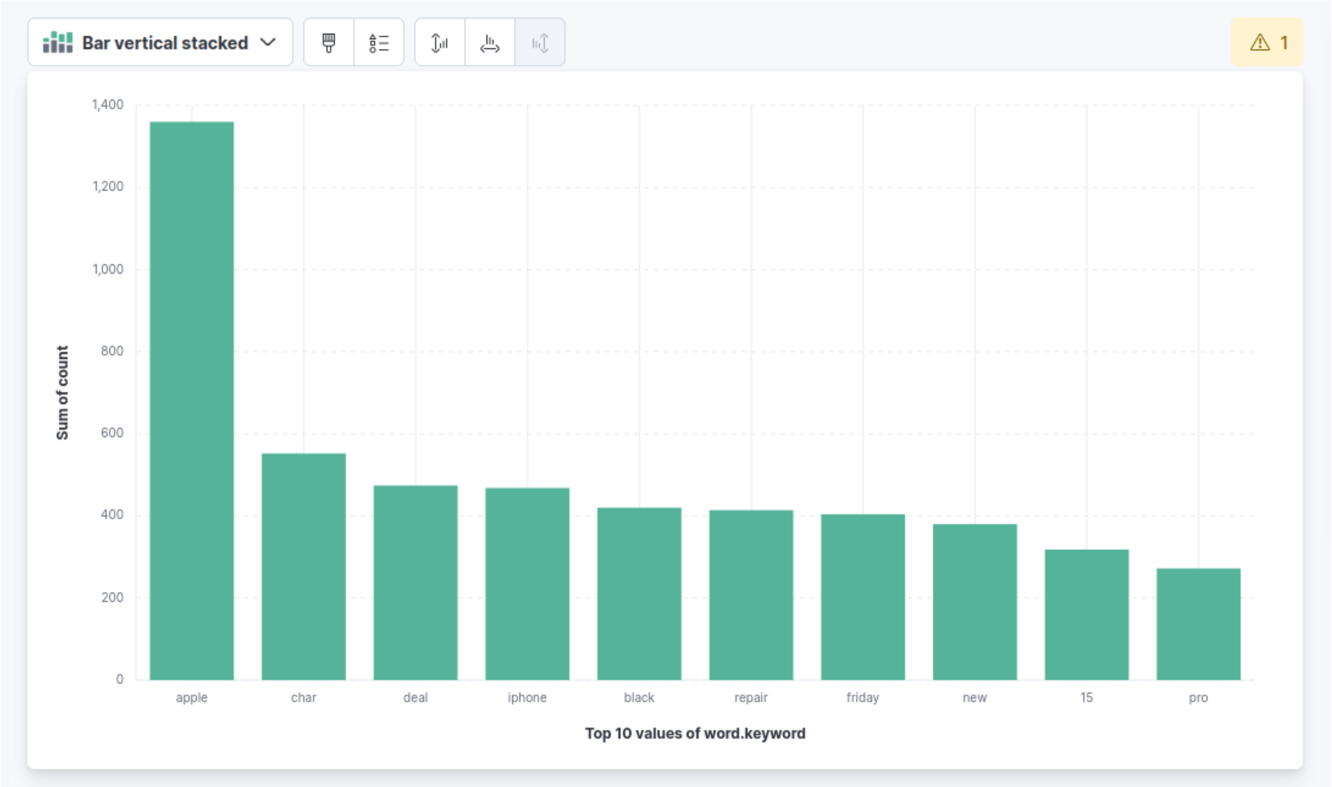

Figure 1: Named Entity Counts after 5+ minutes

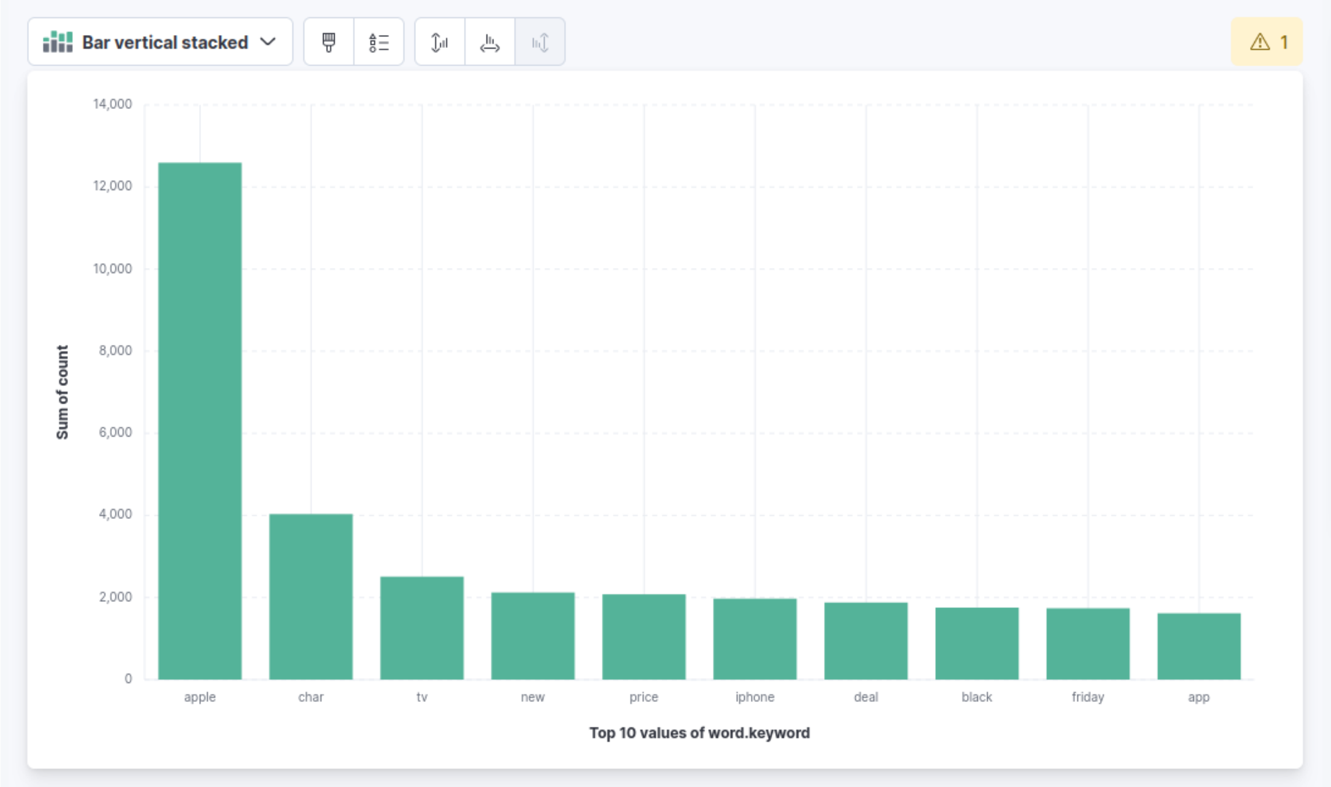

Figure 2: Named Entity Counts After 15+ minutes

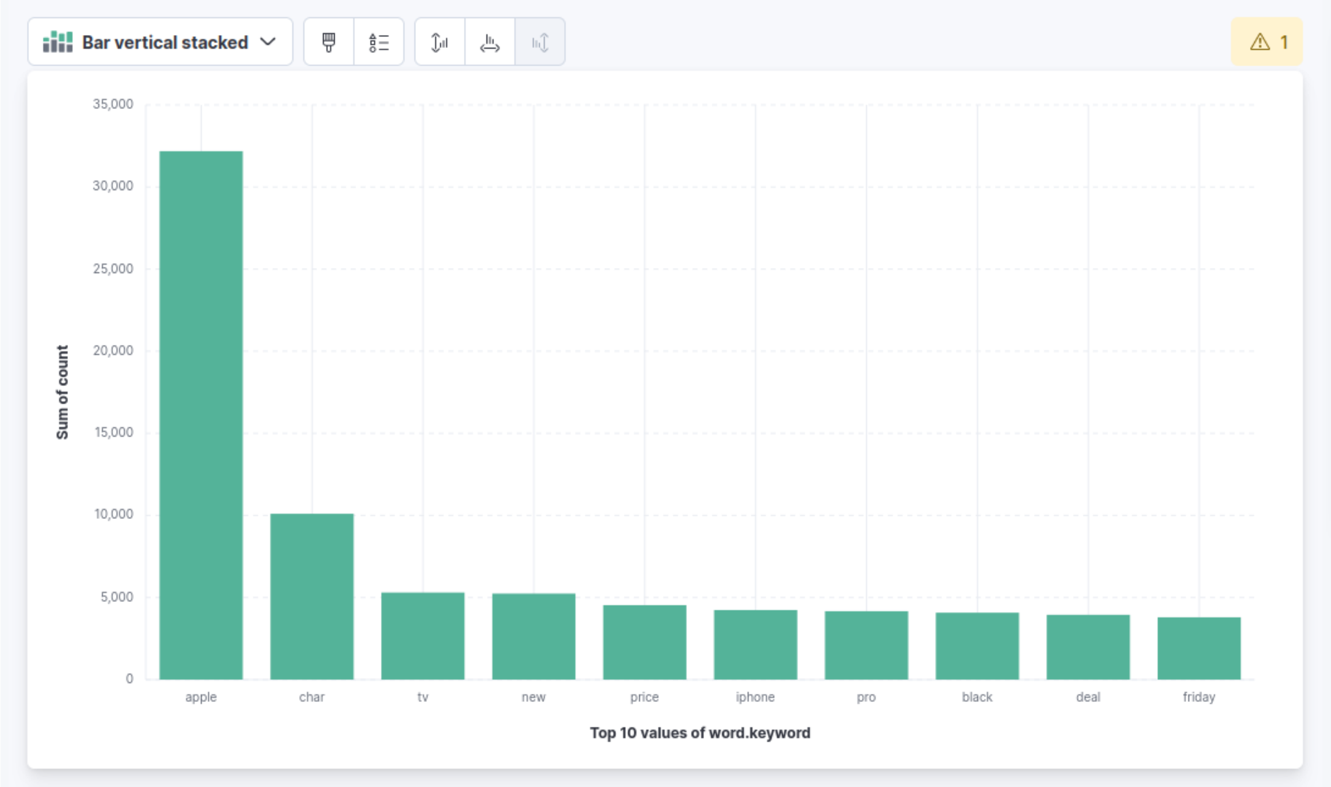

Figure 3: Named Entity Counts after 25+ minutes