Jasen Jones, Miki Seltzer, Eric Whyne
W205 Fall 2015

W205 Final Project Proposal

Previous Research
According to a 2013 Twitter study, Twitter users are more likely than the general population to shop on both Black Friday and Cyber Monday (though the difference is much smaller on Black Friday). Twitter released another report in 2014 aimed at helping companies understand the shopping patterns of Twitter users on Black Friday and Cyber Monday, noting that Twitter users were about twice as likely to pursue electronics deals. Analyzing tweets, therefore, has the potential to reveal important comparisons and consumer sentiments on two of the largest sales days of the year.

Due to the major marketing investment many companies make for Black Friday and Cyber Monday, it's unsurprising that research already exists into what companies and products are the most popular among consumers. Several studies, such as this study into Kohl's Twitter metrics, take an in-depth look at tweets about an individual retailer, tracking sentiment and interest over time. A more robust example from 2014 aggregates twitter data from Black Friday and Cyber Monday and assesses the frequency of hashtags and company names, while also performing basic sentiment analysis. This kind of study provides an example of how we envision our final deliverable, essentially monitoring Twitter traffic and providing aggregated results.

Previous public research, however, seems to focus on Twitter traffic on the actual date of Black Friday and Cyber Monday rather than the preceding days and weeks. We see an opportunity to provide some unique analysis by monitoring Twitter traffic between and in advance of these dates. This may be particularly relevant as an increasing number of companies have deals leaked weeks in advance, which may propagate in some Twitter channels.

Our Proposal
We plan to monitor Twitter traffic in the weeks leading up to the holiday kickoff, through the weekend and possibly even a short while after Cyber Monday. We believe that there may be interesting insights to pull from looking at different aspects of Black Friday and Cyber Monday tweet patterns. We plan to focus on consumer electronics brands, since it is the second most purchased category of products on Black Friday (behind clothing). In particular we plan to focus on large brands like Samsung, Apple, Sony, Lenovo, Dell, Canon, Nikon, Microsoft, Vizio, etc.

Some of the questions we hope to answer by analyzing this Twitter data are as follows:

- What relationship does the volume of tweets during the period leading up to Black Friday/Cyber Monday have with the volume of tweets on and between Black Friday/Cyber Monday?
- What electronics brands are most popular on Black Friday and Cyber Monday?
- Is Black Friday more popular than Cyber Monday as judged by tweet volume?

- Which brands over- or underperformed when looking at tweet volume leading up to Black Friday/Cyber Monday?
- How popular are other “named” shopping days with electronics brands? For example, Grey Thursday (Thanksgiving day, shopping starts before it is officially Black Friday) or Small Business Saturday (the day after Black Friday)?
- Which companies have the best and worst sentiment on Black Friday/Cyber Monday (this may be a little ambitious for the time restraints of this project)?

In order to categorize tweets, we plan to use a list of keywords or hashtags, and collect only tweets that are of interest to us. Given the massive volume of tweets that will inevitably be available during the holiday kickoff weekend, we will need to be judicious about how we choose to filter our data. One likely scenario is to require that each tweet that we collect contain the name of the brand we are interested in, not just general Black Friday and Cyber Monday tweets. It may also be possible to create a more comprehensive list with each brand's major product lines (iPhone, iPad for Apple, Surface, XBox for Microsoft, etc.). We may also need to do research on whether brands have special holiday hashtags that we can take advantage of.

Tools

Trawler is a Python library and set of utilities that allow scraping of Twitter through the Twitter API and in accordance with it’s rate limit requirements. It uses Twython to abstract API access and builds on this by handling Oauth keys and ensuring scrapes don’t result in bans from the Twitter API. This tool will be used to initiate keyword based twitter scrapes during the schedule and dump the resulting tweets to json formatted files on disk. The files will be one-record-per-line to ensure that splitting and ingestion into big data technologies and parallelization can take place with minimal effort.

After scraping, the next step of the tool chain will be the use of hadoop utilities to copy the files into a data lake distributed file system, hdfs. This will be accomplished in one of two ways, either via hadoop -copyFromLocal utility or via Amazon Web Services Elastic Mapreduce directly from an S3 object store. If multiple systems were used to conduct scraping, they will be managed via GNU Parallel and parallel-ssh (pssh). This will allow us to send commands to multiple systems and efficiently manage operations prior to and during the transition of the files from the scraping environment to the data lake analysis environment.

After files are pulled into data lake, pyspark will be used for the subsequent transforms, counting, and analysis. The first step being to filter the data down to only the data necessary for our analysis through the use of filter commands. By conducting filters early in the tool chain we will increase the efficiency of later operations. After the data has been reduced to manageable size and parallel aggregations are done, we will write the output as text to the data lake, and then download for more nuanced statistical analysis and visualization using R Studio. R Studio will allow us to view detailed information about potential correlations and to make observations about statistical significance of findings.