# Starting the Scraper

The command to start the scraper is as follows:

scrapy crawl *spidername*

Currently, these are the spider names, and their respective download source:

|  |  |
| --- | --- |
| googleplay | Google Play |
| apktop | APKTOP |
| fdroid | F-Droid |
| teamapk | Team APK |

**Example:**

To scrape from Google Play, the following command would be used:

scrapy crawl googleplay

# Settings

The scraper settings are defined in the file settings.py.

## Item Pipelines

Information scraped by Scrapy is represented by an item. The ITEM\_PIPELINES setting defines each pipeline, or “transformation”, an item goes through, before that item has been completely scraped. These pipelines are defined in the file pipelines.py. If a new transformation needs to be applied to an item, then a new pipeline should be implemented.

### SQLiteStorePipeline

Stores the information (name, version, developer, etc.) for each APK file in the SQLite database.

### GooglePlayDownloadPipeline

Uses the third-party Google Play API to download APK files from the Google Play store.

### APKFilesPipeline

Extends FilesPipeline to download APK files from sources other than Google Play, and names them with their unique identifier.

## Downloader Middlewares

Downloader middlewares are specified in middlewares.py.

### RandomUserAgentMiddleware

Randomly uses one of the user agents in USER\_AGENT\_LIST in settings.py to scrape.

### ProxyMiddleware

Randomly uses one of the proxies defined in PROXY\_LIST in settings.py at random to scrape.

## Other Settings

Documentation on the remaining settings, and more further documentation on the topics above can be found [here](http://doc.scrapy.org/en/latest/topics/settings.html).