Northwestern MSDS-459 Knowledge Engineering

Assignment 2: Preparing Data for the Knowledge Graph

Andrew Stevens

May 1, 2022

(1) Abstract

Collecting data to build a digital representation can be a daunting task. A company has may facets, internally and externally which we will attempt to catalog and use to construct into knowledge base in graph form. The current stage of efforts involves data preparation for insertion into the database. The subject of this study Tesla has competitors which we will collect data against and determine what additional steps are necessary. Now that the initial scraper has been built and we have an idea of what data will be returned, we must develop the process for making it useful.

(2) Introduction

Since the Model S, Tesla’s first vehicle, began production in 2012 the company has grown considerably (Stringham, 2012). The lineup has grown by 3 vehicles, able to compete with other car manufactures on more platforms – which also makes its graph more complex. Using the guidelines for Porter’s five forces (Dobbs, 2014) reveal areas which will not be as easily scraped for development and which will require more manual work such as customers, and perhaps less so suppliers. After the current stage of engineering data from raw scraping into through natural language processing and into/attachments to entities in a graph, a review of next steps will be required.

(3) Literature review

Ernst & Young (2017) study showed how the rising tide of electric vehicle says to that point (and even since) has lifted all boats (EV manufacturers) as market penetration increases and EVs substitute more of the internal combustion share. Although Hydrogen Fuel-Cell is the most viable automobile technology substitute for EV, my attempts thus far to fetch useful data or research has been limited as it is not currently a market-viable competitor (Manoharan, 2019), but must still at least be manually acknowledged in this study.

(4) Methods

The Wikipedia-focused crawler (Miller, 2022) provided a starting point for fetching industry and company data from Wikipedia articles. The crawler was initialized from the Tesla Inc page (https://en.wikipedia.org/wiki/Tesla,_Inc), fetched all href's from within that page, and the DEPTH_LIMIT parameter was set to one due to the high number of links within the single page (many references involved in the company's history and technology).

A second crawler was developed using the list of competitors fetched from autosaur (2015) for a start_urls list, and DEPTH_LIMIT parameter was set to two as an initial setting of one did not include many useful pages, to be further discussed in results. The spacy and nltk packages were integrated for processing of the language within the articles, and recognition and labeling of entities for use in the graph. The pipelines module of scrapy was enhanced to integrate the processing of text by nltk/spacy, and will be further expanded to yield a more focused result prior to inclusion in the graph.

(5) Results

The results of each of the crawlers with an adjustment that was required for the competitors' is displayed in table 1

Table 1: Scraping Statistics

	item_scraped_count	time	Response_count
wikipedia	1302	22 min	1361
Competitor initial	3022	26 min	4620

Competitor - filtered	3202	35 min	4337

The Wikipedia spider had a much higher “success” rate as it was restricted to the “en.wikipedia.org” domain. The competitor spider crawled unrestricted and was blocked by many as the parameter “ROBOTSTXT_OBEY = True” ensures “friendly” scraping. When following auto manufacture website link trails, many ended in foreign domains, i.e. “.fr” for France. As these pages have foreign languages that would not be helpful in our graph as it will be restricted to English language, two stages were introduced to make the crawling more efficient and focused. The “pycountry” package was installed to easily recognize common country codes in URLs, and “langid” was engaged to classify a page’s language by attempting to translate. Non-english URLs & text were excluded from the crawl steps of either requesting or saving (the former being more efficient, the latter ensuring throroughness).

The resulting two json files contain 4504 lines of data ready for processing and inserting into a database. I have begun including another pipeline step in the crawler which will make calls to the postgres database using the sqlalchemy package in order to insert each record into the target table as it is processed (building an efficient pipeline if periodic crawling was necessary/beneficial for a future project). Two layers of data have been saved – the crawl results and the crawl process as metadata. I will create a secondary metagraph from the crawl results since link parents were stored through a “meta={'parent'” argument within the Request item.

Vehicle model details and direct competitors could be built out using a resource such as JD Power (2022), which includes a “Top Alternatives” component ripe for scraping. A more difficult aspect of competitive analysis would include Federal Incentives and other government influence (Zhou, 2014). Since Tesla crossed the threshold of selling 200,000 vehicles in 2019, the United States federal incentives phase out, giving a more established company a financial disadvantage. A company that has reached his point should have reached economy of scale and can remain profitable, as Tesla has, but it is still an advantage provided to competitors or new entrants (Munzel, 2019).

(6) Conclusions

The graph now has a rough foundation to be built out, with companies and some products. More industry information will need to be curated and fetched from more structured sources such as wikidata. This will allow for a more solid backbone to be carefully constructed as an ontology with some aspect of taxonomy to structure the nodes and their relationships.

(7) References

1. Stringham, Edward Peter, et al. “Overcoming Barriers to Entry in an Established Industry: Tesla Motors.” California Management Review, vol. 57, no. 4, Aug. 2015, pp. 85–103. DOI.org (Crossref), https://doi.org/10.1525/cmr.2015.57.4.85.

2. E. Dobbs, Michael. “Guidelines for Applying Porter’s Five Forces Framework: A Set of Industry Analysis Templates.” Competitiveness Review, vol. 24, no. 1, Jan. 2014, pp. 32–45. DOI.org (Crossref), https://doi.org/10.1108/CR-06-2013-0059.

3. Reed, Eric. “History of Tesla: Timeline and Facts.” TheStreet, https://www.thestreet.com/technology/history-of-tesla-15088992. Accessed 2 May 2022.

4. “EXPANDING THE ELECTRIC VEHICLE MARKET IN U.S. CITIES — An Ernst & Young Report.” International Council on Clean Transportation, July 2017.

5. “Msds-459-2022-Spring/Wikipedia-Crawler at Main · CIGraphs/Msds-459-2022-Spring.” GitHub, https://github.com/CIGraphs/msds-459-2022-spring. Accessed 2 May 2022.

6. Manoharan, Yogesh, et al. “Hydrogen Fuel Cell Vehicles; Current Status and Future Prospect.” Applied Sciences, vol. 9, no. 11, June 2019, p. 2296. DOI.org (Crossref), https://doi.org/10.3390/app9112296.

7. “Car Brands: A Complete and Updated List.” Autosaur, 23 June 2015, https://www.autosaur.com/car-brands-complete-list/.

8. adsf

9. 2022 Tesla Model 3 Ratings, Pricing, Reviews and Awards | J.D. Power. https://www.jdpower.com/cars/2022/tesla/model-3. Accessed 2 May 2022.

10. Zhou, Yan, et al. “Plug-in Electric Vehicle Market Penetration and Incentives: A Global Review.” Mitigation and Adaptation Strategies for Global Change, vol. 20, no. 5, June 2015, pp. 777–95. DOI.org (Crossref), https://doi.org/10.1007/s11027-014-9611-2.

11. Blanco, Sebastian. “Tesla Warns Federal Tax Credit Expires In 5 Weeks.” Forbes, https://www.forbes.com/sites/sebastianblanco/2019/11/30/tesla-warns-federal-tax-credit-expiring-in-5-weeks/. Accessed 2 May 2022.

12. Münzel, Christiane, et al. “How Large Is the Effect of Financial Incentives on Electric Vehicle Sales? – A Global Review and European Analysis.” Energy Economics, vol. 84, Oct. 2019, p. 104493. DOI.org (Crossref), https://doi.org/10.1016/j.eneco.2019.104493.

13.