Northwestern MSDS-459 Knowledge Engineering

Assignment 1: Data Organization and Collection Plan

Andrew Stevens

April 18, 2022

(1) Abstract

Knowledge graphs can be catalog and contextualize data about our piece of the world, expediting and augmenting our ability to exploit them. A car manufacturer, Tesla, is the area of interest which will be the targeted in this graph construction effort – as a graph can provide perspective within a company's scope and the domain surrounding it. In order to construct the graph, we will scrape data from the web on the target company, its suppliers and customers, and its competitors. The entities comprising the graph will adapt as the study progresses, through supervised and unsupervised selection and pruning.

(2) Introduction

A company's products, services, market, and competitors compose a complex architecture which can be represented in a knowledge graph (Martin, et al. 2021). We will study Tesla and its place in the competitive landscape of car manufacturing. Companies enter their industries as either competitors or substitutes, and Tesla Motors could be considered the latter. Tesla "produces a top-selling luxury car and has a market capitalization twice that of Fiat Chrysler and half that of General Motors or Ford" (Stringham, et al. 2015). As the (now) industry leading Electric Vehicle (EV) manufacturer, Tesla must consider competitors, potential substitutes, and its customer/manufacturing bases.

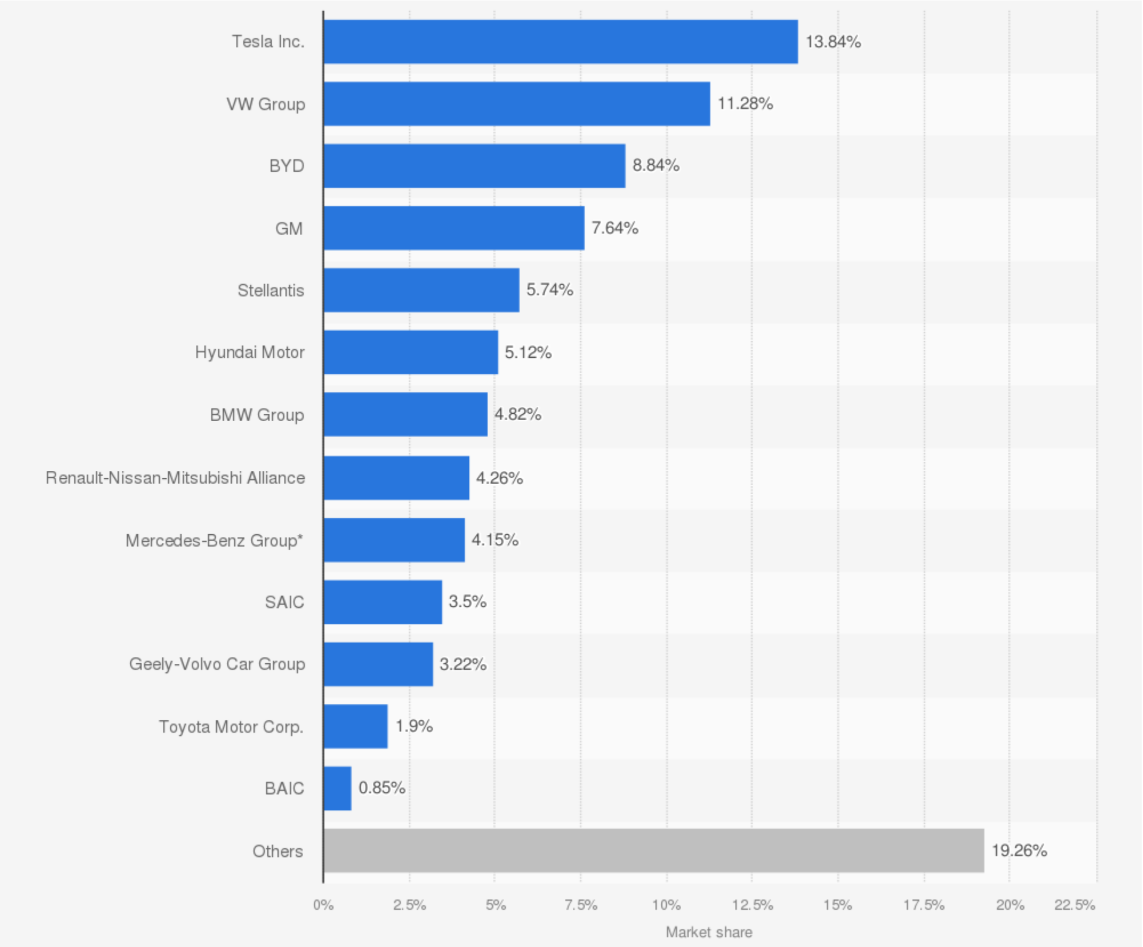

Figure 1: Electric Vehicle Market Share in 2021, by main producer (Statista, 2021)

To assemble Tesla's knowledge graph, we will scrape websites providing information surrounding each of these areas.

(3) Literature review

Chakrabarti provides a thorough layout of how one may approach the retrieval of data required for a graph surrounding a company (1999). This will serve as the main guideline for the collection/ stage of this study. Cheng-Lin-Li recorded an academic approach to automated knowledge graph construction through web crawling and information extraction (2018) and supporting market trend prediction. Thornton (2020) provides some useful code and methodology for pythonically approaching the effort, though the implementation will be different because a focus restricted to Wikipedia is not sufficient for our needs.

(4) Methods

The first stage of scraping will begin with a list of 97 automotive manufacturers retrieved from Autosaur (2015). A scraper will be constructed to traverse starting with 2 levels deep to obtain more industry information. The following initial fields will be recorded for each website (and may be expanded as the effort progresses):

- url – the Uniform Resource Locator of the page retrieved/scraped
- node_type – recognize which nodes are 3rd – level entities strictly for gathering vs manufacturer (a node_type) represented in the automotive domain
- crawl_ts – epoch time site was crawled
- crawl_depth – level of crawl page was found (0 represents a page manually added, 1 a link on a level 1 page and 2 a link found on a level 1 page…)
- crawl_parent – ID of page a level higher which led to this link (retain crawl path)
- Relationships – to make connections to other nodes within the graph

The pages will be processed through NLP for entity extraction, testing of whether entity recognition will be necessary, potential topic modeling for use in relevance and training of classifer for future pages. A distiller will be developed to filter & refine content of pages scraped for inclusion in the graph & review of graph sufficiency.

(5) Results

The Autosaur article (2015) was initially used to recognize other manufacturers in the industry and collect the first round of metadata and relationships about and amongst these nodes.

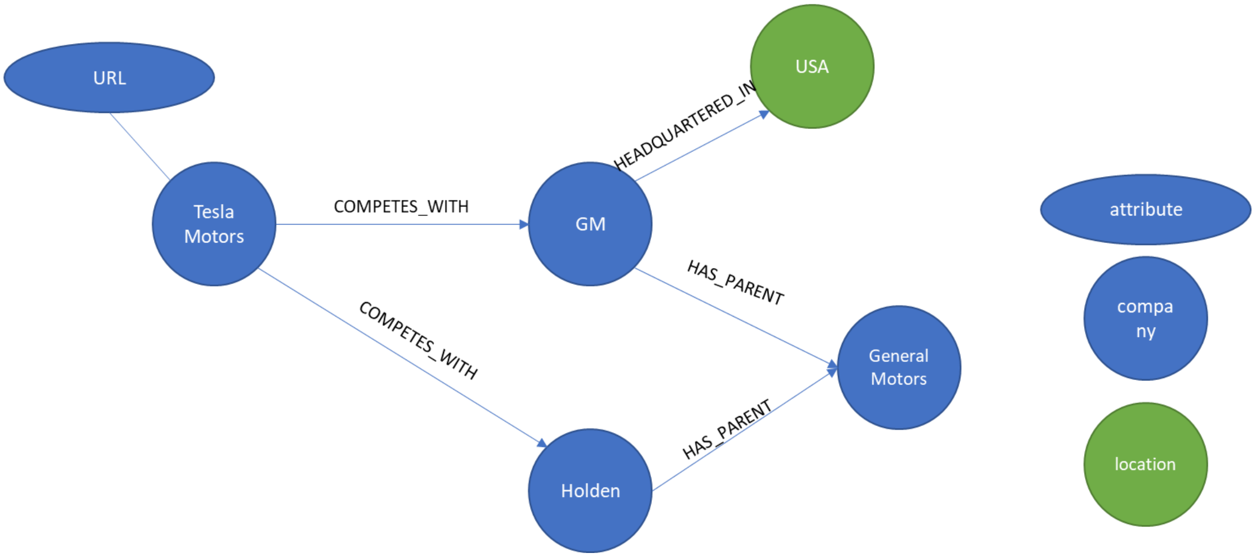

Figure 2: Sample Graph with initial nodes and relationships

This provided the initial set of nodes and relationships to be recognized:

- Companies will be related to other companies as parents or competitors
- Locations will be connected to companies as where they are located
- Attributes such as URL will be stored within a node's metadata

Parsing the page allowed the extraction of each of the companies' names, where they are headquartered, whether they have a parent company, and these are used to create an initial set of connections. The URLs stored for each of these nodes will be scraped, classified and distilled for further use.

(6) Conclusions

The automotive industry is extensive and limiting the crawl to the most relevant information will be a significant effort. Multipipe data engineering, statistics and machine learning methods will be necessary to build a useful knowledge graph.

(7) References

1. Martin, Sean, et al. The Rise of the Knowledge Graph. 2021.

2. 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.

3. EV-Volumes.com. (March 7, 2022). Global plug-in electric vehicle market share in 2021, by main producer [Graph]. In Statista. Retrieved April 18, 2022, from https://www.statista.com/statistics/541390/global-sales-of-plug-in-electric-vehicle-manufacturers/

4. Cheng-Lin-Li (2018) Knowledge Graph [Source code]. Accessed Retrieved April 18, 2022 from https://github.com/Cheng-Lin-Li/KnowledgeGraph.

5. Thornton, Chris. “Auto-Generated Knowledge Graphs.” Medium, 2 Oct. 2020, https://towardsdatascience.com/auto-generated-knowledge-graphs-92ca99a81121.

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

7. “Training Pipelines & Models · SpaCy Usage Documentation.” Training Pipelines & Models, https://spacy.io/usage/training/. Accessed 18 Apr. 2022.

8. Chakrabarti, Soumen. 2003. Mining the Web: Discovering Knowledge from Hypertext Data. New York: Morgan Kaufman. [ISBN-10: 1-55860-754-4] Chapter 2. Crawling the Web (pages 18–43).

9. Chakrabarti, Soumen, Martin van den Berg, and Byron Dom. 1999, May 17. Focused crawling: a new approach to topic-specific Web resource discovery. Computer Networks: The International Journal of Computer and Telecommunications Networking, 31(11-16): 1623–1640.

10. Miller, Thomas. W. 2015. *Web and Network Data Science: Modeling Techniques in Predictive Analytics.* Upper Saddle River, N.J.: Pearson. [ISBN-13: 978-0-13-388644-3].

11. Olston, Christopher, and Marc Najork. 2010. "Web Crawling." *Foundations and Trends in Information Retrieval,* 4(3): 175–246.