Knowledge Graph Task

Introduction:

The Task is to extract information about Elon Musk and present this information in a graph.

As requested in the task we need to extract information about Elon Musk from raw data text, I assumed that the raw data is already provided from scrapping or from the text the notebook starts with the text provided about Elon musk obtained from Wikipedia.

Data:

The data is assumed to be scraped or provided as text.

Tools:

* Spacy → to extract entities pairs and relations
* Regex → to clean the data
* NetworkX → for presenting the graph

Process:

The process started by splitting the data into sentences (tokenized), then passed throw functions the extract pairs of entities from each sentence and the relation between these entities these functions utilize spacy POS and NER models to extract them.

Regex was used to coreference the pronouns (he mainly) I tried to use the coreference model from Huggingface but couldn't install it.

Then we formed pandas data frame contains the entities and the relation

Then passed the data frame to the NetworkX with proper arguments to return a graph of Elon musk relations.

Challenges:

The first challenge was utilizing spacy to extract the relations and paired entities in the text, I had to go throw multiple articles and examples to get it done.

The second challenge was coreference the pronoun to its origin.

Learned lessons:

Spacy is a very powerful tool that could achieve a lot in information extracting tasks.

Very useful models like coreference models.

References:

<https://towardsdatascience.com/from-text-to-knowledge-the-information-extraction-pipeline-b65e7e30273e>

<https://neptune.ai/blog/web-scraping-and-knowledge-graphs-machine-learning>

<https://www.analyticsvidhya.com/blog/2019/10/how-to-build-knowledge-graph-text-using-spacy/>

<https://github.com/varun196/knowledge_graph_from_unstructured_text>