Natural Language Processing

Knowledge Graph On Wikipedia Sentences for Information Extraction

Initial Review Document

Team Members

Kriti Gupta - 17BCE1327

Amrit Gupta - 17BCE1082

Project Introduction

Wikipedia is a minefield of information and in most cases the first resource that is referred to when in search of factoid or in answer to any question. It is the first resource that pops up in most searches and therefore the first resource most people click on in order to find answers to their questions. But more often than not the articles on the subjects are long and convoluted and thus finding what is needed becomes a task in itself. This is the problem we hope to address. We plan to scrape wikipedia documents to create the required corpus and develop a knowledge graph based on it. On this we further plan to create a question answering mechanism that would provide answers to specific questions.

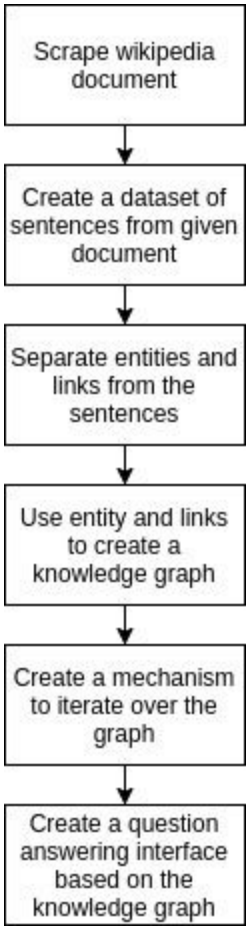
Project Design

Softwares used:

Python v3 libraries

- Pandas
- Re
- Spacy
- Networkx
- Matplotlib
- bs4

Project flow:



Methodology

To start with, a dataset relevant to the problem is needed and hence we plan to scrap the web for sentences from wikipedia documents to create a knowledge graph on. For the creation of the knowledge graph itself a couple of steps will be required. Firstly, the corpus will need to be cleaned and preprocessed in the form of stemming, lemmatization and removal of stopwords. Next, we need to extract the entities from the sentences in the form of parts of speech tags such as nouns, pronouns, etc. This is to be accomplished through the spaCy library in python. Next we need to extract the relations to form the links of the knowledge graph. This is done by finding the root, ie, the verb in the sentences using the same library. In the previous steps, entity pair relations are maintained where the subject and object of the sentences are identified along with the root or verb as the link from subject to object. Finally these will be put into a knowledge graph form. A method of traversal for this graph will be defined maybe through the use of the library node2vec in python.

Dataset details

The dataset shall be composed of sentences from a wikipedia document .This shall be formed on our own, rather than taken to increase authenticity of domain. A sample can be seen below:

2	confused and frustrated, connie decides to leave on her own.
3	later, a woman's scream is heard in the distance.
4	christian is then paralyzed by an elder.
5	the temple is set on fire.
6	outside, the cult wails with him.
7	it's a parable of a woman's religious awakening—
8	c. mackenzie, and craig vincent joined the cast.
9	later, craig di francia and action bronson were revealed to have joined the cast.
10	sebastian maniscalco and paul ben-victor were later revealed as being part of the cast.
11	we just tried to make the film.
12	we went through all these tests and things
13	m global was also circling to bid for the film's international sales rights.
14	canadian musician robbie robertson supervised the soundtrack.
15	it features both original and existing music tracks.
16	it is the worst reviewed film in the franchise.
17	but she injures quicksilver and accidentally kills mystique before flying away.
18	military forces tasked with her arrest.
19	the train is attacked by vuk and her d'bari forces.
20	kota eberhardt portrays telepath selene gallio,
21	singer did not return to direct the sequel, x-men:
22	the last stand, which was written by penn and simon kinberg.
23	jessica chastain was also potentially being considered for the same character.
24	mauro fiore served as cinematographer.
25	filming was completed on october 14, 2017.
26	the soundtrack was released digitally on june 7.
27	the album was released digitally on august 2, 2019.
28	the film is distributed by walt disney studios motion pictures.

Tentative Team Contributions:

Both the team members will contribute equally in the coding and creation of the knowledge graph. The Individual demarcations may be:

Amrit: Creation of Dataset

Integration of knowledge graph

Question answer code and UI

Kriti: Preprocessing of dataset

Identification of entities and links

Creation of entity pairs

Graph Traversal Code