CS6320: NLP PROJECT REPORT

TEAM NAME: TRANSFORMERS

Team members: -

Raghav Murali – RXM190067

Venkatesh Sankar – VXS200014

Problem Description

The purpose of this project is to implement a Question Answering System using NLP features and techniques on SQuAD dataset. **S**tanford **Qu**estion **A**nswering **D**ataset (SQuAD) is a reading comprehension dataset, consisting of questions posed by crowdworkers on a set of Wikipedia articles, where the answer to every question is a segment of text, or *span*, from the corresponding reading passage, or the question might be unanswerable.

Our goal is to implement an end-to-end pipeline that accepts a question sentence as input, analyzes a set of input articles, and returns the most probable answer sentence to the question along with the article ID the sentence was taken from.

The question types are of the following patterns: -

1. WHAT questions: -
   Examples: What act was repulsive to Romans?
   What company did Ray own?
2. WHEN questions: -
   Examples: When was the invasion of Gaul by Rome?
   When did Apple go public?
3. WHO questions: -
   Examples: Who founded Apple Inc.?
   To whom was John married?

Proposed Solution

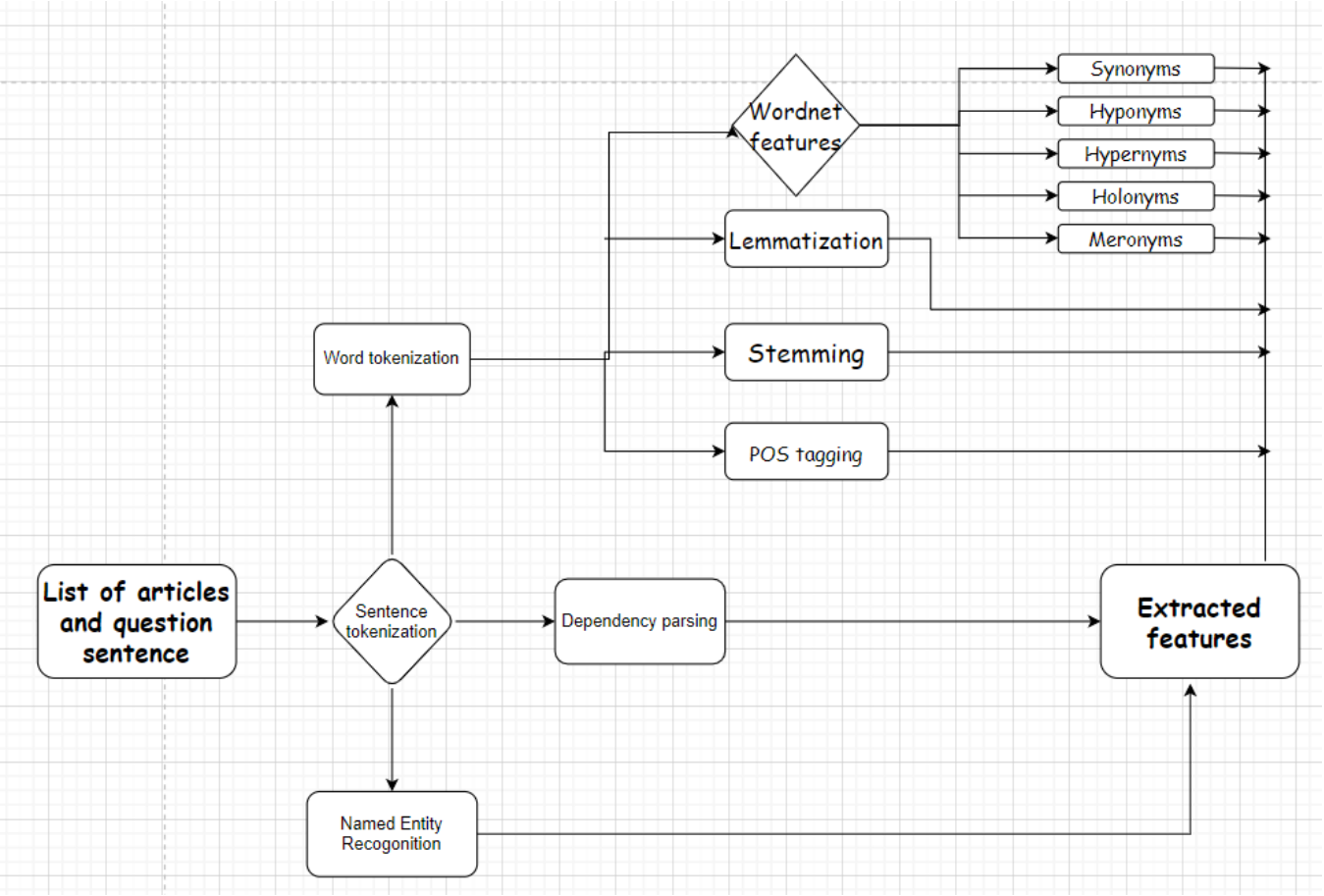
To build a successful NLP pipeline, it is important to understand the input question and the articles. Using various python packages such as NLTK and Spacy, we will be preprocessing the input data and extract features which will be useful to answer the questions. The feature extraction used techniques such as tokenization, lemmatization, part-of-speech tagging, wordnet features etc. Further, we will be using Elastic Search to index the input data and query the indexed data with the question sentence. We then use Cosine Similarity, Named Entity Recognition and dependency parsing to correctly identify the answer to the question.

Tools Used

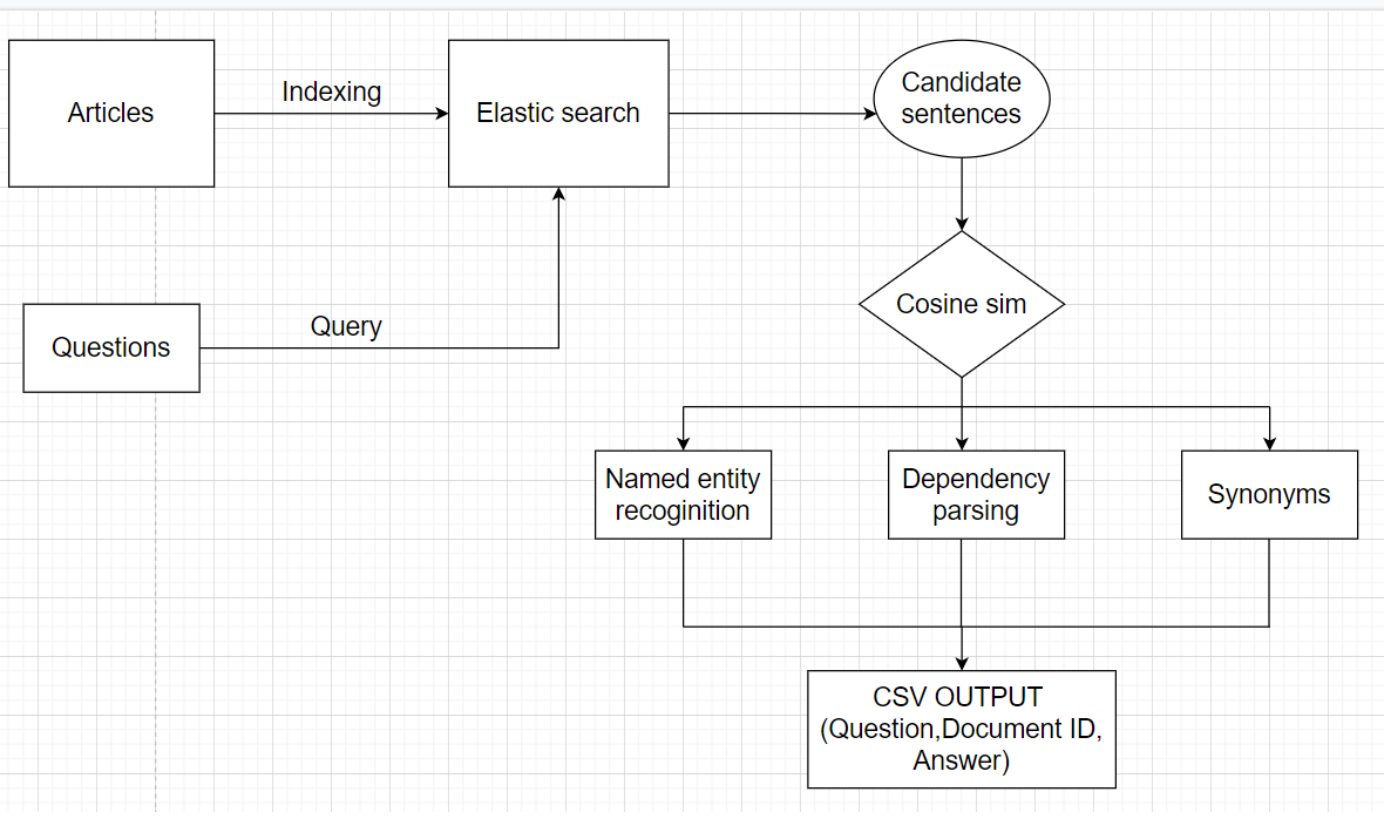
Name	Description
Python	Python is an interpreted high-level general-purpose programming language. Its design philosophy emphasizes code readability with its use of significant indentation. Its language constructs as well as its object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects
NLTK	**NLTK (Natural Language Toolkit)** Library is a suite that contains libraries and programs for statistical language processing. It is one of the most powerful NLP libraries, which contains packages to make machines understand human language and reply to it with an appropriate response.
Spacy	**spaCy** is a free open-source library for Natural Language Processing in Python. It features NER, POS tagging, dependency parsing, word vectors and more.
Elastic Search	Elasticsearch is a search engine based on the Lucene library. It provides a distributed, multitenant-capable full-text search engine with an HTTP web interface and schema-free JSON documents.

Architecture diagram

Task 1 – Feature Extraction



Task 2 – NLP Pipeline implementation



TASK 1 – FEATURE EXTRACTION

For the task 1 in the project, we focused on extraction of a list of NLP based features from the articles in the dataset and natural language questions.

1. Word tokenization -> word_tokenize(text_data)
   Purpose: Splits a sentence into words using NLTK library.

2. Sentence tokenization -> sent_tokenize(text_data)
   Purpose: Converts text in the articles into individual sentences.

3. Stemming and Lemmatization-> word_stemmatization(words), lemmatization(word_tokens)
   Purpose: Converting the words into their root forms.

4. POS tagging -> pos_taggers(word_tokens)
   Purpose: Mapping each word with their POS tag.

5. Wordnet features -> wordnet_features(word_tokens)
   Purpose: Extract Synonyms, hypernyms, hyponyms, meronyms and holonyms from the articles.

6. Dependency parsing -> dependency_parsing(sentence)
   Purpose: Extract word dependency tags from sentences and output a parse tree using Spacy's inbuilt dependency parser.

7. Named Entity Recognition -> named_entity_recognition(sentence)
   Purpose: Use Spacy's inbuilt NER functionality to assign labels to named entities within a sentence.

TASK 2 – NLP PIPELINE IMPLEMENTATION

1. Elastic Search
   By using elastic search, the articles are indexed with individual sentences and mapped with their corresponding article ids and sentence numbers.
   Once the articles are indexed, the search query is run on the question data after removing stop words and the candidate sentences matching the search query are returned in JSON format.

2. Cosine Similarity
   After the elastic search results are gathered, the cosine similarity score is calculated between the candidate sentences and the question using NERs, dependency parsing and synonymy.

3. Named Entity Recognition (NER)
   As per the 3 question formats, a list of expected NERs is created as follows and used as the filtering criteria for the next step: -
   WHO – Person, Organization
   WHEN – Date, Time
   WHAT – Person, Organization
   Depending on the NER match between the question and the answer sentences, a score is appended to the cosine similarity scores to set the precedence order.

4. Dependency parsing
   By using the inbuilt spacy dependency parser, we are extracting the root, head, and the dependency parse tree. By comparing the root of the question with the root of the candidate answer sentences, we are appending a score to the sentences from the previous step depending on the results from the root matching.

5. Synonymy
   By using the Wordnet synonymy feature, we are matching all the possible synonyms from the question sentence and checking for a match in the candidate answer sentences and assigning a score if there is a match.

TASK 3 – IMPLEMENTATION AND RESULTS

Input format:

1. List of articles with the SQUAD Dataset
2. A text file containing a list of questions (one per line)

Output format:

A CSV file containing the following: -

1. Input question
2. Supporting article ID.
3. Supporting sentence containing the answer to that question.

Output from sample data: -

1. Question -> Who mediated the truce with Khomeini?
   Answer sentence -> Subsequently, Khomeini accepted a truce mediated by the UN.
   Document ID -> 400.txt
   Main tool used to derive answer -> Named Entity Recognition (PERSON)

2. Question -> When did an empire collapse after Alexander's conquests?
   Answer sentence -> The empire collapsed in 330 BC following the conquests of Alexander the Great.
   Document ID -> 400.txt
   Main tool used to derive answer -> Named Entity Recognition (TIME)

3. Question -> What is the nickname for Tucson?
   Answer sentence -> Roughly 150 Tucson companies are involved in the design and manufacture of optics and optoelectronics systems, earning Tucson the nickname Optics Valley.
   Document ID -> 390.txt
   Main tool used to derive answer -> Cosine similarity

4. Question -> Who sold Arizona?
   Answer sentence -> Arizona, south of the Gila River was legally bought from Mexico in the Gadsden Purchase on June 8, 1854.
   Document ID -> 390.txt
   Main tool used to derive answer -> Dependency parsing

5. Question -> When was Arizona purchased by Mexico?
   Answer sentence -> Arizona, south of the Gila River was legally bought from Mexico in the Gadsden Purchase on June 8, 1854.
   Document ID -> 390.txt
   Main tool used to derive answer -> Synonyms

6. Question -> What distance can the Fajr-3 missile travel?
   Answer sentence -> The Fajr-3 (MIRV) is currently Iran's most advanced ballistic missile; it is a liquid fuel missile with an undisclosed range which was developed and produced domestically.
   Document ID -> 400.txt
   Main tool used to derive answer -> Synonyms

Output sample: -

	question	document	Answer Sentence
2	Who mediated the truce with Khomeini?	400.txt	Subsequently, Khomeini accepted a truce mediated by the UN.
3	When did an empire collapse after Alexander's conquests?	400.txt	The empire collapsed in 330 BC following the conquests of Alexander the Great.
4	What is the Leader of the Revolution also known as in Iran?	400.txt	Iran has also developed a biotechnology, nanotechnology, and pharmaceuticals industry.
5	What is the nickname for Tucson?	390.txt	Roughly 150 Tucson companies are involved in the design and manufacture of optics and optoelectronics systems, earning Tucson the nickname Optics Valley.
6	Who sold Arizona?	390.txt	Arizona, south of the Gila River was legally bought from Mexico in the Gadsden Purchase on June 8, 1854.
7	When was Arizona purchased by Mexico?	390.txt	Arizona, south of the Gila River was legally bought from Mexico in the Gadsden Purchase on June 8, 1854.
8	What type of fuel is used by Fajr-3 missile?	400.txt	The Fajr-3 (MIRV) is currently Iran's most advanced ballistic missile, it is a liquid fuel missile with an undisclosed range which was developed and produced domestically.
9	Who succeeded Reza Shah?	400.txt	In 1935, Reza Shah requested the international community to refer to the country by its native name, Iran.
10	What led to students capturing the US embassy?	400.txt	On November 4, 1979, a group of students seized the United States Embassy and took the embassy with 52 personnel and citizens hostage, after the United States refused to return M
11	Who is the Supreme Leader?	400.txt	It has not challenged any of the Supreme Leader's decisions.
12	What distance can the Fajr-3 missile travel?	400.txt	The Fajr-3 (MIRV) is currently Iran's most advanced ballistic missile, it is a liquid fuel missile with an undisclosed range which was developed and produced domestically.

Issues Encountered

1. Some of the answer sentences were not part of the initial result returned from Elastic search, due to which the query had to be tweaked.
2. Spacy's NER filtering incorrectly removes some of the named entities which were useful in finding the answer to the question.

Pending issues

1. In certain cases, the whole compound terms cannot be identified through the current level of dependency parsing.
2. Some of the correct answers are not retrieved from the Pipeline as the number one candidate sentence and are located further down the precedence order.

Potential Improvements

1. For rule-based approach when it comes to NERs, more accurate NER filters for WHO, WHEN and WHAT questions may yield better results and cover corner scenarios.
2. Adding features such as coreference resolution and custom NERs may improve the model.