Question Answer System

Bruno Zecchi

This paper explains the question-answer system developed to provide answers to three types of questions. The system uses concepts like sentence tokenization, NER, cosine similarity, and TF-IDF to arrive at a final set of possible answers to the question. The process starts off by getting user input on the type of question and components of the question. It then extracts documents relevant to the question and runs analysis on the documents' sentences. Those final sentences are further analyzed, and a final set of answers is derived from them. A more detailed explanation of this process can be seen below. In the next sections there are also outputs of sample questions for each of the three types of questions. Main conclusions are that the system does a good job of answering the questions "which companies went bankrupt on month X of year Y?" (Question 1) and "who is the CEO of company X?" (Question 3). This ease might be because sentences that contain many of the keywords for both questions are answering the question very well. A sentence that has a company name, the word "CEO", and a person's name predicts very well that the person is the company's name. Likewise, a sentence that has a given month and year, the word "bankrupt", and organization names predicts very well that those organizations went bankrupt in that time period. The system has a harder time with the question "what affects GDP?" (Question 2). An explanation for this drop in performance is that, unlike "Person Name" and "Organization Name", factors of GDP do not have a unique entity type. We can very easily have sentences that contain "GDP", "rise" or "drop", and a percentage, but it is much harder to identify what word refers to the thing that is causing the GDP to rise/drop by that percentage. For that reason, sample output contains all nouns in top sentences. Additionally, if there are multiple percentages in the sentences, it hard to identify which is the change in GDP. Thus, all percentages in same sentence(s) as selected noun are displayed as possible changes. Future improvements of the system would make it possible to infer this causality relationship in sentences and extract only relevant nouns.

Analysis Process

Preprocessing: Build TF-IDF table of all tokens in all documents.

Step 1: Get question type and inputs from user. From this query, select keywords. The relevant entity type for the answer is thus set.

- **Question 1**
  - Keywords: Bankrupt, Bankruptcy, Month X, Year Y
  - Entity Type: Organization
- **Question 2**
  - Keywords: GDP, Rise, Increase, Drop, Decrease, Effect
  - Entity Type: Percentage
- **Question 3**
  - Keywords: CEO, Company X
  - Entity Type: Person

Step 2: Documents that have at least one keyword in them are extracted. This is to reduce processing time for the next steps.
Step 3: All of the extracted documents are given a score for how close they are to the questions' keywords. The scoring process is described below:

a) For each document, the TF-IDF of all keywords are extracted.
b) Any document that has a score of 0 for any keyword is discarded.
c) The median for all keywords' TF-IDF value is calculated.
d) Subset documents that have keywords' value above median for all keywords. This is to ensure that we select documents that have wide variety of all keywords, and not just many of one.
e) Sum TF-IDF scores for all keywords for all subset documents. Choose documents with highest sum of values. For Question 3, only select top document. For other questions, select top 10 documents.

Step 4: Concatenate top document(s) into one string and split into sentences.
Step 5: Apply NER on sentences and extract sentences that have relevant entity type. Only these sentences will be used for further analysis.
Step 6: Concatenate query into the list of sentences and run cosine similarity on them. Select sentence(s) that have the highest similarity to the query row. For Question 3, only select top sentence. For Question 1 and Question 2, select top 10 sentences.
Step 7: For Question 1 and 3, extract the words that are the matching entity type. For Question 2, extract all nouns and percentages from the sentences. Connect the two types of words with each other to know which percentage is related to which noun.
Step 8: Present words as final possible answers.

Sample Output

Question 1

- Which companies went bankrupt in February of 2008?

```
Q1: Which companies went bankrupt in month X of year Y?
Q2: What affects GDP?
Q3: Who is the CEO of company X?

Select Q1, Q2, or Q3: Q1

Month X: February

Year Y: 2008
Possible answers:
{'Congress', 'Daily Mail', 'Hollande', 'eBay', 'FOMC', 'Lehman', 'EFH',
'Lehman Brothers', 'Energy Future Holdings Corp.'}
```

- Which companies went bankrupt in June of 2004?

```
Month X: June

Year Y: 2004
Possible answers:
{'LightSquared', 'the European Court of Human Rights', 'Buiter to Mees',
'OGX', 'Columbia', 'Skype', 'Yukos', 'SecondMarket', 'Assured Guaranty
Municipal Corp.', 'Trump Organization'}
```

- Which companies went bankrupt in November of 1999?

```
Month X: November

Year Y: 1999
Possible answers:
{'aÂ\xa0', 'Global Inc', 'Reuters', 'Segway', 'American Airlines', 'Ponzi',
'the U.S. Bankruptcy Court', 'Merrill', "Trump Entertainment's", 'U.S.
Bankruptcy Court'}
```

Question 2:

- What affects GDP? – Yuan. What percentage of drop or increase is associated with this property?

```
Possible answers:
['headwinds', 'zone', 'charts', 'impact', 'yuan', 'appreciate', '1.5%.Â',
'outlook', 'housing', 'reason', 'supply', 'bargain', 'countries',
'management', '#', 'consensus', 'budget', 'questions', 'today', 'view',
'sequester', 'account', 'ones', 'rates', 'recovery', 'weeks', 'spending',
'US.Â', 'revenue', 'addition', 'cuts', 'deal', '0.6%.Â', 'taxes', 'trade',
'stocks', 'surplus', 'state', '1st', 'economy', 'consumption', 'release',
'year', 'part', 'market', 'posts', 'money', 'stock', 'dollar', 'quarter',
'level', 'government', '+0.1', 'affect', 'defense', 'policy', 'jobs',
'recession', 'exports', 'forecasts', 'deficit', 'euro', 'country', 'chance',
'tomorrow', 'regions', 'terms', 'share']

Select factor to see percentage change in GDP associated with factor: yuan
{'approximately .6%'}
```

- What affects GDP? – Consumption. What percentage of drop or increase is associated with this property?

```
Select factor to see percentage change in GDP associated with factor:
consumption
{'22%'}
```

- What affects GDP? – Housing. What percentage of drop or increase is associated with this property?

```
Select factor to see percentage change in GDP associated with factor: housing
{'above 3% (4%'}
```

Question 3:

- Who is the CEO of Apple?

```
Company X: apple
Possible answers:
{'Tim Cook', 'Carl Icahn'}
```

- Who is the CEO of Tesla?

```
Company X: Tesla
Possible answers:
{'Hop', 'Musk'}
```

- Who is the CEO of Facebook?

```
Company X: Facebook
Possible answers:
{'Mark Zuckerberg'}
```