IEMS Homework 4

Q & A

Zi Mei

Executive summary:

Natural Language Processing is a study area that focuses on interactions between computers and human natural languages, making computers could understand basic human languages. NLP has been widely used from simple machine translation to advanced question and answer system, especially when virtual online assistance applications such as chat box, are growing fast, I believe the need for Q&A will booming in future both in text and voice input requests.

I’m given the two years of articles from Business Insider to develop a simple Q&A system that be able to answer questions like ‘Who is the CEO of company X?’, ‘Which companies went bankrupt in month X of year Y?’ and ‘What affects GDP? What percentage of drop or increase is associated with this property?’. This is a rather simplified Q&A system with simple query input interface and the answer’s accuracy level is not always satisfying. However, more sophisticated Q&A system could be created based on this and apply to Business Insider website to help online users browse news and find their content more easily. For instance, users could input single questions they are interested, and pull out specific answers efficiently. Also, new users could use online assistant based on this Q&A model to help them navigate the whole website more conveniently.

This Q&A system could also help website accumulate data on users’ most frequently asked questions or most interested topics, thus, websites could better understand user’s preference and create more attractive articles and contents.

Problem statement:

I’m given two years article from Business Insider and asked to develop a simple Q&A system that is capable of answering questions like ‘Who is the CEO of company X?’, ‘Which companies went bankrupt in month X of year Y?’ and ‘What affects GDP? What percentage of drop or increase is associated with this property?’.

Assumption:

While employing tf-idf scores, I assume the more frequently some key terms are mentioned in one document while less mentioned in other documents, the more likely the terms are important.

For GDP question, I assume it is more likely a noun will lead to the GDP’s fluctuation and the more it mentioned, the more possible this noun is the answer.

Methodology:

Python was used for this development, and packages like nltk are installed. For indexing the corpus, I used inverse index method, extracting all the related articles with target words and rank them by tf-idf score. The computer indexed the whole corpus after finish scanning the whole text data.

For the search part, main function here is to use the index to retrieve relevant document according to input queries. First, tokenize the input question and extract the key words, pull out all the articles with key words and sort them based on tf-idf scores, and at the end, computer would return all the related articles.

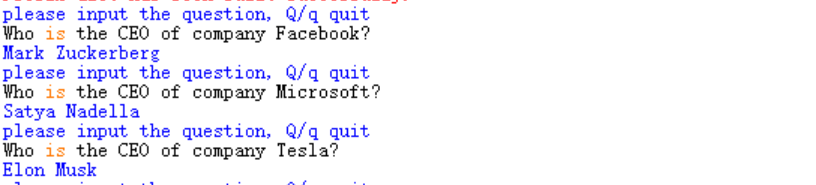
To analyze input question, I need to classify question type so the system will put input query to the right type. To achieve this function, I classified the question with key word for each type. For company went bankrupt question, the key word is ‘bankrupt’; for CEO question, the key word is ‘CEO’; for GDP question, the key word is ‘GDP’ and ‘percentage’.

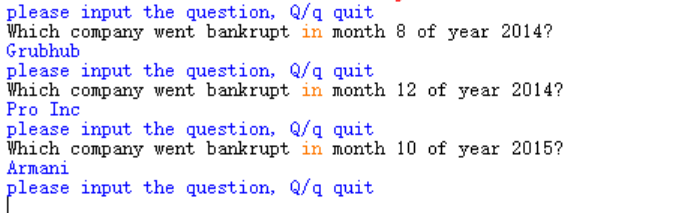
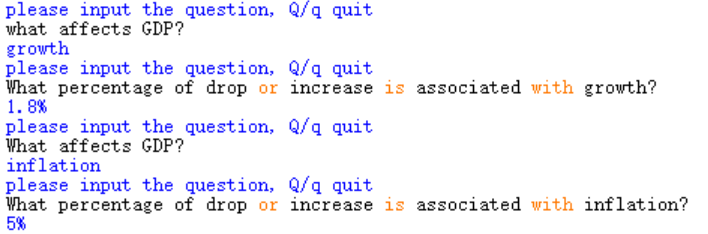
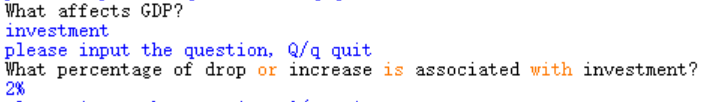
When input the three types of questions, the system will search the final answer within the returned relevant documents from the previous step. For question which company went bankrupt, the system first searches the index with keywords ‘bankrupt’ ‘month’ and ‘year’ and filter out the unrelated information that not mentioned in query and return the candidate sentences with highest score. Then extract the company name with highest frequency.

For question what affects GDP and what the percentage associated with that factor, first, search the index with target words like ‘GDP’, ‘percent’ and ‘percentage’ and etc. and return the most related documents. Then I assume the nouns that could be more possibly leads to the GDP changes, and then system will put more weight on noun words parts and return the candidate answer with the highest score. And for the following question, the system will extract the associated number that closest to the entity.

For CEO of which company question, the method is similar to find the answer of company went bankrupt one. The system search through the most relevant articles via the indext with key words like ‘CEO’, and return the candidate answer which is closest to key word CEO and has highest appearance frequency.

Analysis:



Instruction to use:

Make sure the computer install Python succesfully

Please run all the codes and install the packages listed at the source code

Wait untill computer finish and the screen will appear instructions like ‘please input Question, Q/q quit.’

Then input the question like ‘who is the CEO of company Tesla?’, wait until the answer appears, and this could take several seconds. And the answer should be Elon Musk.

Conclusion:

The Q&A system contains basic functions and is eligible to find the three types of questions and return reliable answers. It could be used to extract informations corresponding to input queries.

Next step:

Although the simplied Q&A system is able to search and return candidate answers based on input questions, it is limited to answering only these three particular questions. The next step could be diversifying the input question type that could be answered by this system.

It is also meaningful try to improve the answers’ accuracy and thus improve the whole system’s quality. If more time given, maybe bigger text data could be used to improve the system’s quality.