CS410 – Course Project

Topic Modeling on Quora Q & A

Sushanta Panda (panda5@Illinois.edu)

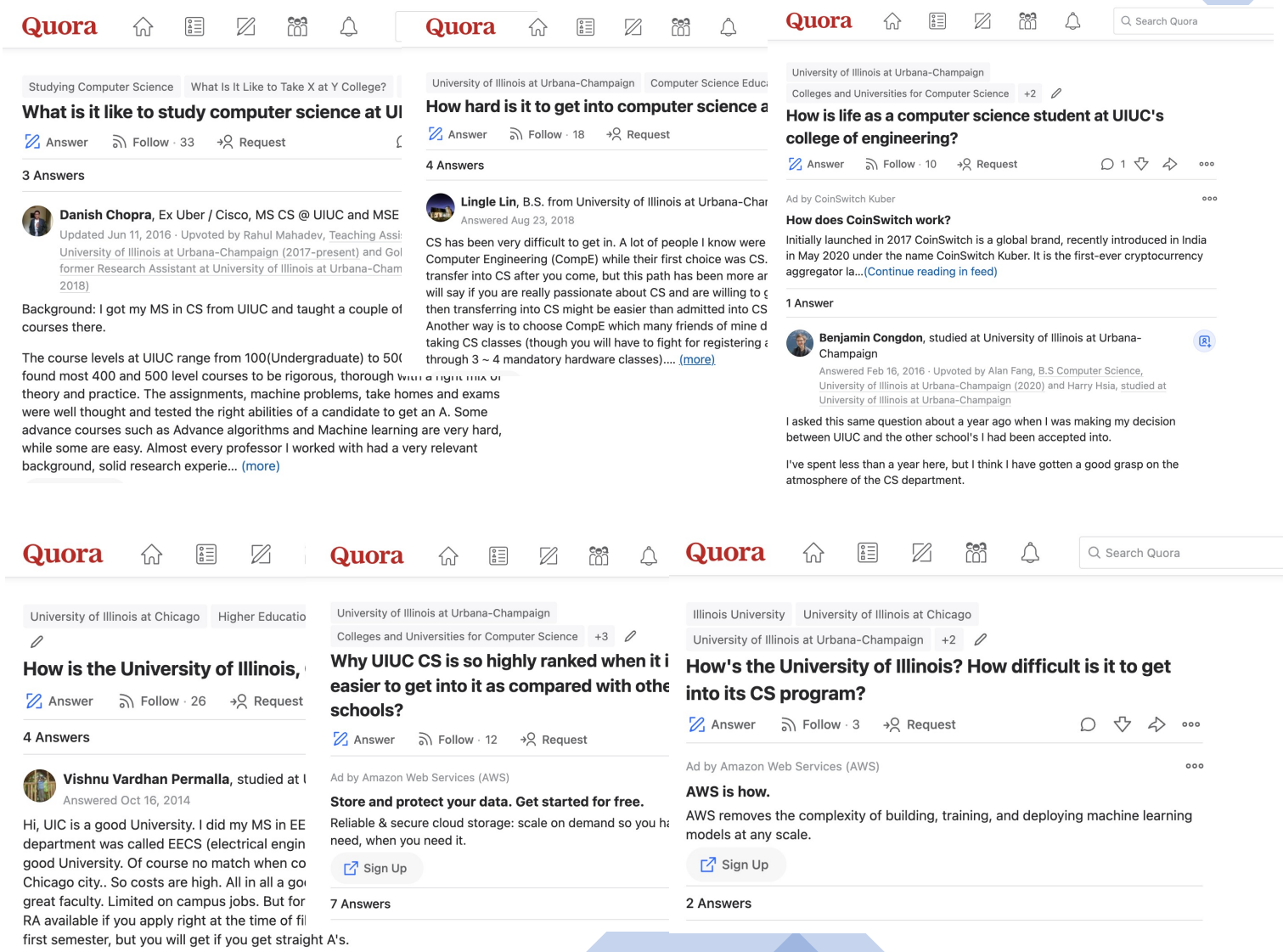
Introduction

- Many students try to explore various universities before they finally enroll into one of the University. All of them explore / search the internet to know more about the university, it's curriculum, research areas, student's academia and other interests. This search / re-search is somehow a tedious process and consumes lots of time. This Project is an attempt to simply the process and gives students the To 5 / 10 Topics discuss in the internet users discussed about certain university. This gives a glimpse of certain ideas what the university (or certain department of a university) is all about. For e.g., in the Top topics are discussed about "research" / "Professor" / "Good" / "Excellent", then it creates an impression that the University or the searched department of a University is very well performed, as folks are discussing about the "research ideas" , and "Good" / "Excellent" are the key topics mentioned in most of the documents. In other case, if the top topics discussed around "Tiring" / "Long" / "bad" etc., it doesn't give an idea the certain department is doing well.

- One can certainly be not 100% sure from the top topics, however it gives an ample idea what are the top things users are discussing about. Since internet is so broad and big, we limit ourself to the quora (www.quora.com) platform as the Q & A and limit the department to CS (Computer Science) and University is University of Illinois. Though we can configure the entire pipeline to pulled all these information from different platform (other than quora) and for other universities and other department, we are currently focus on how this prototype project will really work, rather to scale it up to certain levels.

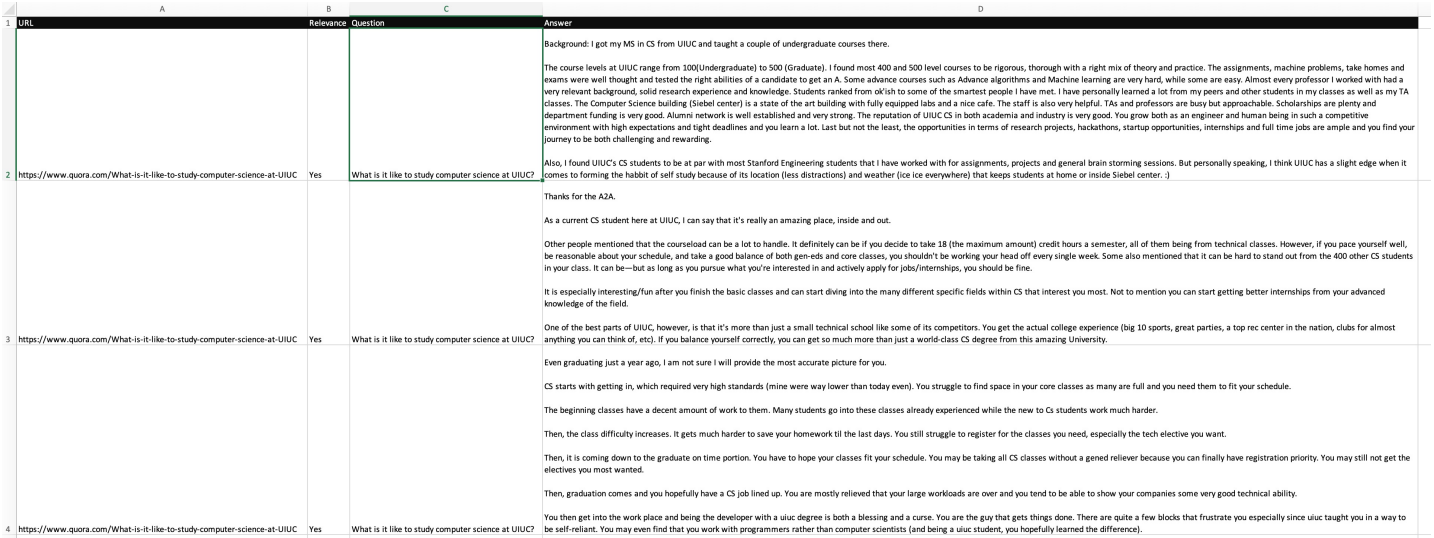
Datasets

As part of the dataset, we have search in the google about the search words “University of Illinois Urbana Champaign” and “Computer Science” and “quora”, extract all the documents which are available. Based on these documents, we have extracted the ”questions” and all “answers” for each documents. Against one document, one question is attached. Below are some of the documents



Dataset (Contd ..)

- All the questions & answers are being captured in the form of a spreadsheet (manually) to feed into the system for the Topic Modeling. Following are the columns which are stored as part of the data collection
  - **URL** # URL of the quora Question
  - **Relevance** # Whether the question is relevance
  - **Question** # Text content of the question
  - **Answer** # Text content of the Answer



	A	B	C	D
1	URL	Relevance	Question	Answer
2	https://www.quora.com/What-is-it-like-to-study-computer-science-at-UIUC	Yes	What is it like to study computer science at UIUC?	Background: I got my MS in CS from UIUC and taught a couple of undergraduate courses there. The course levels at UIUC range from 100(Undergraduate) to 500 (Graduate). I found most 400 and 500 level courses to be rigorous, thorough with a right mix of theory and practice. The assignments, machine problems, take homes and exams were well thought and tested the right abilities of a candidate to get an A. Some advance courses such as Advance algorithms and Machine learning are very hard, while some are easy. Almost every professor I worked with had a very relevant background, solid research experience and knowledge. Students ranked from ok'ish to some of the smartest people I have met. I have personally learned a lot from my peers and other students in my classes as well as my TA classes. The Computer Science building (Siebel center) is a state of the art building with fully equipped labs and a nice cafe. The staff is also very helpful. TAs and professors are busy but approachable. Scholarships are plenty and department funding is very good. Alumni network is well established and very strong. The reputation of UIUC CS in both academia and industry is very good. You grow both as an engineer and human being in such a competitive environment with high expectations and tight deadlines and you learn a lot. Last but not the least, the opportunities in terms of research projects, hackathons, startup opportunities, internships and full time jobs are ample and you find your journey to be both challenging and rewarding. Also, I found UIUC's CS students to be at par with most Stanford Engineering students that I have worked with for assignments, projects and general brain storming sessions. But personally speaking, I think UIUC has a slight edge when it comes to forming the habbit of self study because of its location (less distractions) and weather (ice ice everywhere) that keeps students at home or inside Siebel center. :)
3	https://www.quora.com/What-is-it-like-to-study-computer-science-at-UIUC	Yes	What is it like to study computer science at UIUC?	Thanks for the A2A. As a current CS student here at UIUC, I can say that it's really an amazing place, inside and out. Other people mentioned that the courseload can be a lot to handle. It definitely can be if you decide to take 18 (the maximum amount) credit hours a semester, all of them being from technical classes. However, if you pace yourself well, be reasonable about your schedule, and take a good balance of both gen-eds and core classes, you shouldn't be working your head off every single week. Some also mentioned that it can be hard to stand out from the 400 other CS students in your class. It can be—but as long as you pursue what you're interested in and actively apply for jobs/internships, you should be fine. It is especially interesting/fun after you finish the basic classes and can start diving into the many different specific fields within CS that interest you most. Not to mention you can start getting better internships from your advanced knowledge of the field. One of the best parts of UIUC, however, is that it's more than just a small technical school like some of its competitors. You get the actual college experience (big 10 sports, great parties, a top rec center in the nation, clubs for almost anything you can think of, etc). If you balance yourself correctly, you can get so much more than just a world-class CS degree from this amazing University.
4	https://www.quora.com/What-is-it-like-to-study-computer-science-at-UIUC	Yes	What is it like to study computer science at UIUC?	Even graduating just a year ago, I am not sure I will provide the most accurate picture for you. CS starts with getting in, which required very high standards (mine were way lower than today even). You struggle to find space in your core classes as many are full and you need them to fit your schedule. The beginning classes have a decent amount of work to them. Many students go into these classes already experienced while the new to Cs students work much harder. Then, the class difficulty increases. It gets much harder to save your homework til the last days. You still struggle to register for the classes you need, especially the tech elective you want. Then, it is coming down to the graduate on time portion. You have to hope your classes fit your schedule. You may be taking all CS classes without a gened reliever because you can finally have registration priority. You may still not get the electives you most wanted. Then, graduation comes and you hopefully have a CS job lined up. You are mostly relieved that your large workloads are over and you tend to be able to show your companies some very good technical ability. You then get into the work place and being the developer with a uiuc degree is both a blessing and a curse. You are the guy that gets things done. There are quite a few blocks that frustrate you especially since uiuc taught you in a way to be self-reliant. You may even find that you work with programmers rather than computer scientists (and being a uiuc student, you hopefully learned the difference).

Code

- From this section onwards will elaborate the details of the code. The code basically has 4 parts as below
  - Import Python Library
  - Importing data (collected manually)
  - Cleaning of the Data
  - Create the model to generate the Top K Topics

Code (Contd…)

- **Import Python Library**

- Below are the some of the python packages needs to import as part of the Program to run
  - pandas
  - Numpy
  - Gensim
  - Nltk
  - Pyldavis
  - Pickle

- Left is the sample of the code

Import Python Library

We have imported all the relevance Library

- Pandas
- Numpy
- gensim
- nltk - Natural Language Tool Kit
- pyLDAvis
- pickle

In [30]:

```
import pandas as pd
import numpy as np
import os
import re

import gensim
from gensim.utils import simple_preprocess
import nltk
nltk.download('stopwords')

import gensim.corpora as corpora
from nltk.corpus import stopwords
from pprint import pprint

import pyLDAvis
import pyLDAvis.gensim_models as gensimvis

import pickle

import warnings
warnings.filterwarnings('ignore')
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     /Users/sushanta/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

Code (Contd…)

- **Importing data (collected manually)**

- The excel sheet which is being captured all the data is converted into the csv file (manually). The csv file is then loaded into a dataframe via “pandas” library. However, since we are interested in only relevance document, will exclude in the next section of “data extraction” process and take out only the “Questions” column.

- Left is the sample of the Code

Import Data

The data is extracted from the www.quora.com. The data is extracted by manually from searching the relevance of UUIC in the Computer Science Department. These documents (text extracted) are stored in the excel sheet (quora_data.xls) with 4 columns as below

- URL
- Relevance
- Question
- Answer

We are interested in those questions and its related answer where "Relevance" = "Yes". These questions are related only to University of Illinois Urbana Champaign (UIUC UC) Computer Science Department, where users are raising question related to know the feedback of the department. Other questions like comparision of of University and other questions, which we have marked as "Relevance" = "No".

In [4]:
```
# Read the Quora data from the quora_data.csv file
quora_data = pd.read_csv('quora_data.csv')

# Print all the Columns
quora_data.columns
```

Out[4]:
```
Index(['URL', 'Relevance', 'Question', 'Answer'], dtype='object')
```

Code (Contd…)

- Cleaning of the Data

- As part of the data cleaning process, we will do the following tasks
  - Removing special characters
  - Removing the stop words. In addition to the nltk's provided stop words, will add few of the stop words which we think are relevance to act as a stop words

Left is the sample code of the Data Cleaning

Data Cleaning

Following Data Cleaning are done as part of the data cleaning

- remove the special character and HTML character (".", "!", "?")
- lower the text data

In [6]:
```
# Remove the special characters (., !, ?, [, ])
quora_data_relevance['Answer_processed'] = quora_data_relevance['Answer'].map(lambda x: re.sub('[,\.!?]', '', x))

# Convert to the lowercase
quora_data_relevance['Answer_processed'] = quora_data_relevance['Answer_processed'].map(lambda x: x.lower())
```

In [9]:
```
stopp_words = stopwords.words('english')

#Below are some of the words which are extended other than included of the english stop words
stopp_words.extend(['uiuc','cs','uic','university','graduate',"illinois",'school',\
                    'also','know','like','get','department','take','many','would',\
                    'chicago','since','taking','go','really','one','may','even',\
                    'lot',"computer","science"])

def convert_words(sentences):
    for sentence in sentences:
        yield(gensim.utils.simple_preprocess(str(sentence), deacc=True))

text_data = quora_data_relevance.Answer_processed.values.tolist() # Convert the text into list
final_data = list(convert_words(text_data))

#Removing the stop words
final_data = [[word for word in simple_preprocess(str(doc)) if word not in stopp_words] for doc in final_data]
```

Code (Contd…)

- Model & Find Top K Topics
- This part of the code creates the model using the LDA and generate the K cluster where each cluster shows the Top Topics associated with the Answers from the Quora

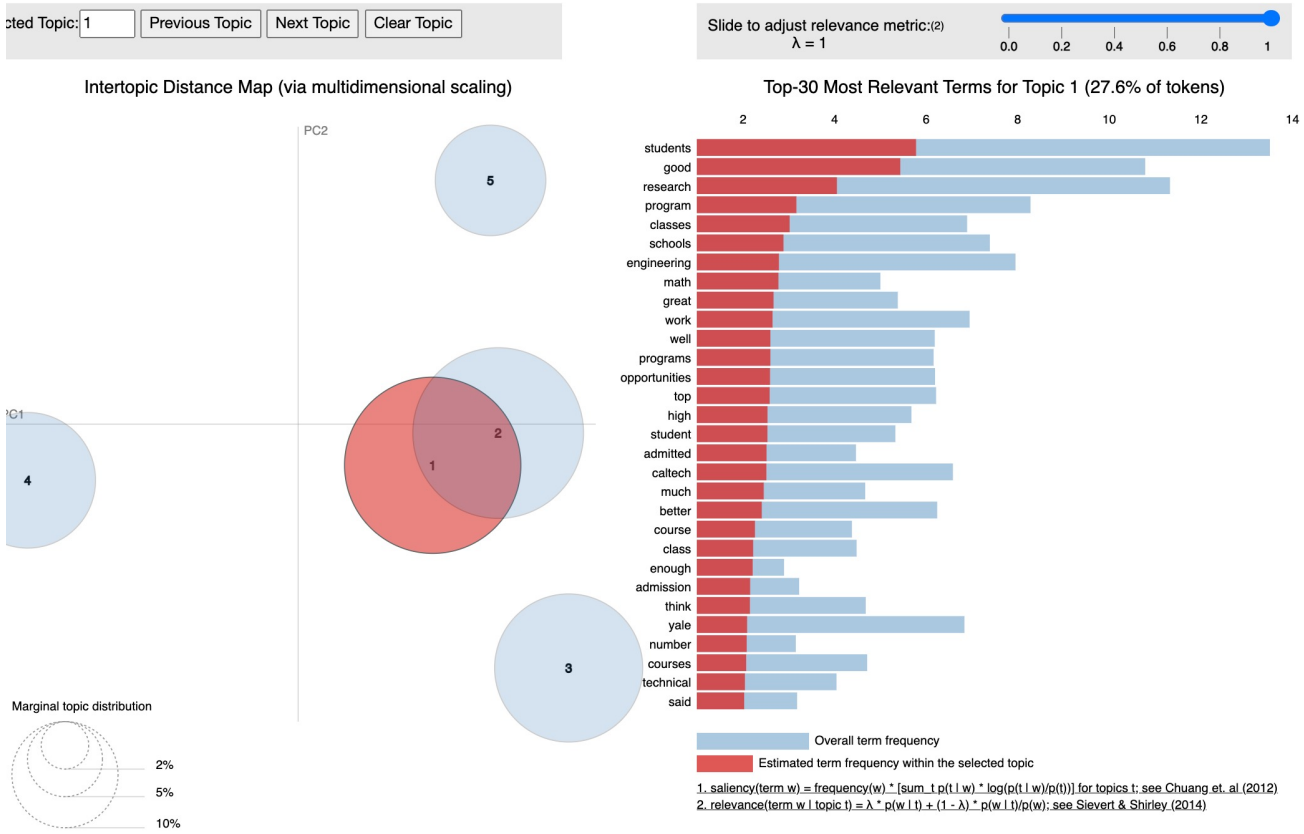
Left is the sample code

In [6]:
```
# Create Dictionary
id_to_word = corpora.Dictionary(final_data)

corpus = [id_to_word.doc2bow(text) for text in final_data]

# Number of Topics
num_topics = 5

# Create the LDA model
lda_model = gensim.models.LdaMulticore(corpus=corpus,id2word=id_to_word,num_topics=num_topics)
doc_lda = lda_model[corpus]
```

In [7]:
```
#Create the folder & file location where the file needs to be dumped
folder_file_location = 'ldavis_'+str(num_topics)
ldavis_file_path = os.path.join(folder_file_location)

# Visualize the topics
pyLDAvis.enable_notebook()

#Create the ldavis model and dump the model in the specified folder
ldavis_model = gensimvis.prepare(lda_model, corpus, id_to_word)
with open(ldavis_file_path, 'wb') as f:
    pickle.dump(ldavis_model, f)

#Save the Model in the HTML form
pyLDAvis.save_html(ldavis_model, folder_file_location +'.html')

ldavis_model
```

Top Topics via Latent Dirichlet Allocation (LDA)

- There are 5 Cluster of Topics are being generated. The 1st Cluster of topics are shown in the image. We can see the top topics are “student” / “good” / “research” / “Program” / “classes”. This means that the UIUC’s computer science department is well received by users and most of them are praising the University and the Computer Science Department field.




The End