MET CS688

Web Analytics and Mining

Zlatko Vasilkoski

Term Project Suggestions

Outline of the CS688 Final Term Projects

Term Project Presentation Dates will be announced .

- **You can also present during the last class**
- **Or before if needed (please arrange this early)**

The duration of each presentation should be at most 15 minutes

R Term Project Suggestions

Here are few term project suggestions, but you can modify these suggestions or choose something else as your term projects:

1. Searching and Ranking a set of given web pages term project
2. Tweeter stock market sentiment analysis term project
3. Sports Data Analytics term project
4. The Web site scrapping term project
5. IIoT forecasting term project
6. TED Talks Text Summarization

R Term Project Suggestions

1) Searching and Ranking a set of given web pages term project. Dataset: Your choice of 10 URLs (suggestion 2-3 sets of similar subject URLs)

- Download the HTML content of these web pages, extract the text content, perform preprocessing and content analysis to rank (hierarchical clustering) these pages for similarity.
- Visually display the 75 most frequent words. Use *gvisBubbleChart* function to visually display the 3 most frequent words (in 3 different colors) in each of the 10 web pages (x axis) while the y axis represents the frequency count of the words. Specify the size of the bubble to represent the word frequency.

2) Tweeter stock market sentiment analysis term project. Dataset: Your choice of 6 stocks, 3 largest gainer and 3 loser stocks for the day. (suggestions: http://finance.yahoo.com/ ; http://www.google.com/finance)

- Use the R Twitter API, to perform stock market sentiment analysis.
- Make necessary addition (if needed) to the lexicons (the positive and negative word lists), and compute the sentiment score of all the tweets for each gainers (losers) set.
- Plot a bar chart of the sentiment. Use *googleVis* R Package to create a candlestick plot of the stock prices for the stocks used for this project and compare them with a chart or table obtained from an online financial source.

Note that a similar kind of analysis (on a large scale) published in 2010 under the title "Twitter mood predicts the stock market" (http://arxiv.org/abs/1010.3003) brought to the authors a multimillion-dollar fortune.

R Term Project Suggestions

3) Sports Data Analytics term project. Suggested R package (library(SportsAnalytics)) & website that you would need to scrape, such as http://www.landofbasketball.com/championships/
Use the SportsAnalytics API for R, to accomplish several sports analytics tasks.

- Retrieve the NBA data for the 2007-2008 season.
- Subset the data for your favorite team. Show the code you used to find:

  Which player has the best three point percentage?

  Which player has played the largest number of minutes?

  Which player has the most "Steals"?
- Show 5 teams for the 2007-2008 season that have the most wins in descending order.
- Use at least 5 Google charts (your choice) to show relevant data from this dataset.
- Use gvisGeoChart function to display the location on the world map all of the Basketball World Cup Champion countries that you can find at:
- http://www.landofbasketball.com/world_cup_stats/medals_by_year.htm

Alternatively, you can choose another sport to perform similar sports analytics tasks.

R Term Project Suggestions

4) The Web site scrapping term project: Dataset: A website of a movie theater of your choice.

- Use the R URL connectivity to a website to access a specific movie theater showtimes of your choice and import the HTML content of the showtimes web page into an R object called “Showtimes”
- Create an R list object called "Program" that will contain all of the movie titles currently playing in the theater.
- Identify and graph all of the links from the R object “Showtimes” to the other web pages.
- Scrape particular text content of your choice (such as review or something similar) from this retrieved web pages and use all the relevant text mining steps.
- Use gvisBubbleChart function to visually display the 3 most frequent words (in 3 different colors) in each of the retrieved web pages. Use pages as x axis while the y axis represents the frequency count of the words. Specify the size of the bubble to represent the word frequency.

5) IIoT forecasting term project. The goal is re-work the IIoT lab project to get the best energy consumption prediction (R-squared) for the month of February 2017 only. Some of the ideas you may want to explore are:

- Think of extracting and flagging seasons and possibly holidays from the time stamp and add them as additional dependent data columns.
- Consider ways to capture floating holidays such as Presidents Day or Washington's Birthday celebrated on the third Monday of February.
- In addition, implement changes in the neural network parameters to optimize the neural network prediction.

R Term Project Suggestions

6) Text Summarization: Dataset: Visit TED Talks website https://www.ted.com/talks and download the text of 10 talks of your choice.

- Download the talks as text files, extract the text content, perform preprocessing if needed, and create corpus from these multiple documents.
- Visually display the most frequent words and perform content analysis to rank (hierarchical clustering) these talks for similarity.
- Use provided Document Summarization code to summarize the text of 10 TED talks of your choice. For each document determine the number of "topics" and the "number of most important sentences" that provide best summarization for each document. Consult the slides on Document Summarization for the meaning of these terms.
- For one document, visually highlight by hand the summary sentences in the full text and save the file in a supporting format of your choice (i.e. MS word, pdf etc.)

- If familiar with Python, use the 3 summarization techniques illustrated in the class code and compare your summaries with the R summarization code.

Python BERT Related Term Project Suggestions

The BERT related Term Projects would carry extra credit (if needed).

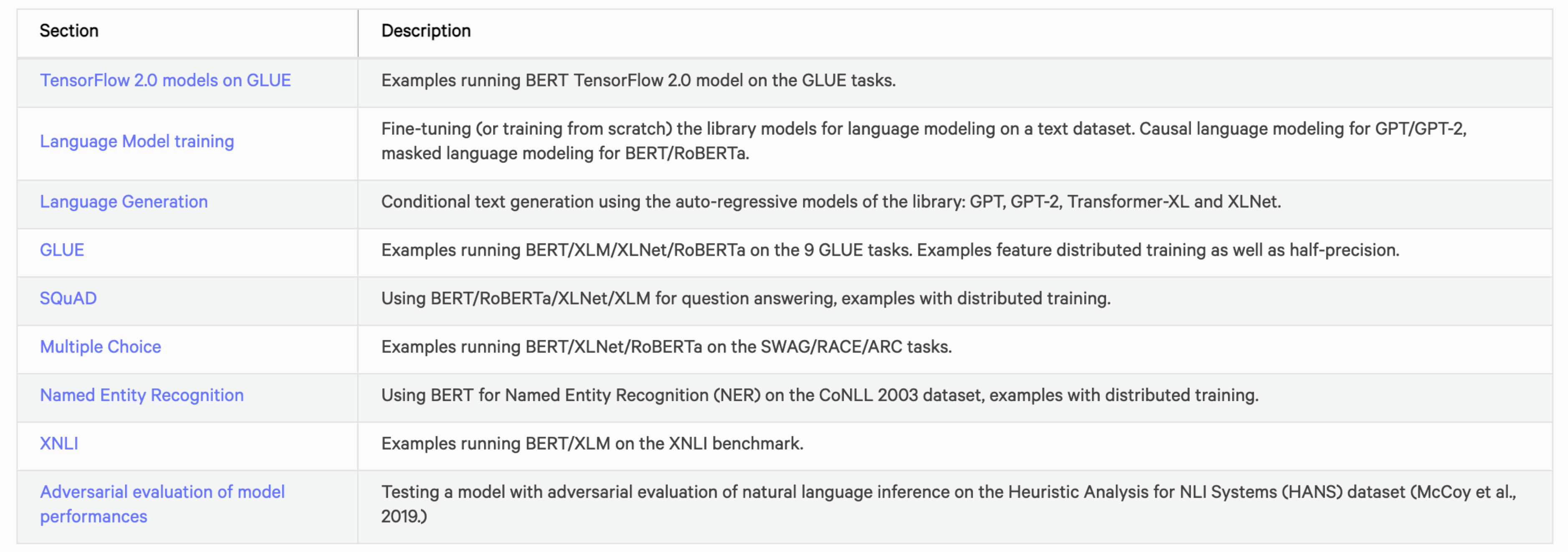
In general, be inventive, come up with tasks that you find interesting and use this opportunity to gain extra credit and more importantly to get introduced to this latest NLP techniques and models which you will find very useful in your future career.

You can use the python starter code on Blackboard which is illustrated on several NLP tasks

- You can use Python & BERT Models for **any** NLP tasks such as Q&A, Sentiment Analysis, Name entity recognition, keyword extraction, language generation etc.

You can also use any online resources such as Huggingface site:

https://huggingface.co/transformers/v2.5.1/examples.html

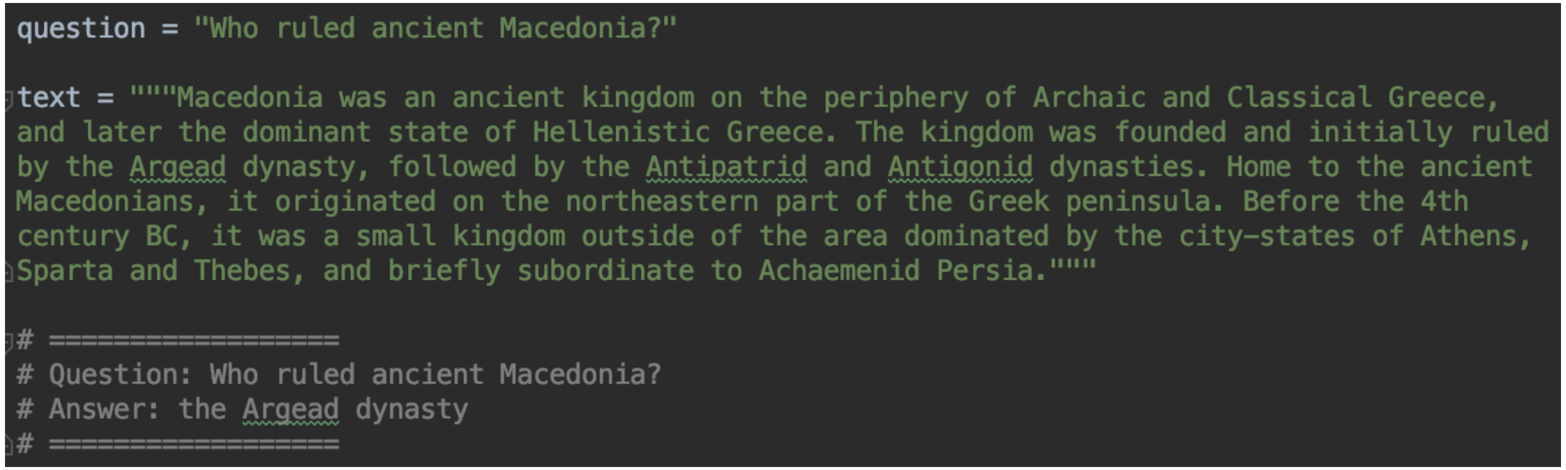


Section	Description
TensorFlow 2.0 models on GLUE	Examples running BERT TensorFlow 2.0 model on the GLUE tasks.
Language Model training	Fine-tuning (or training from scratch) the library models for language modeling on a text dataset. Causal language modeling for GPT/GPT-2, masked language modeling for BERT/RoBERTa.
Language Generation	Conditional text generation using the auto-regressive models of the library: GPT, GPT-2, Transformer-XL and XLNet.
GLUE	Examples running BERT/XLM/XLNet/RoBERTa on the 9 GLUE tasks. Examples feature distributed training as well as half-precision.
SQuAD	Using BERT/RoBERTa/XLNet/XLM for question answering, examples with distributed training.
Multiple Choice	Examples running BERT/XLNet/RoBERTa on the SWAG/RACE/ARC tasks.
Named Entity Recognition	Using BERT for Named Entity Recognition (NER) on the CoNLL 2003 dataset, examples with distributed training.
XNLI	Examples running BERT/XLM on the XNLI benchmark.
Adversarial evaluation of model performances	Testing a model with adversarial evaluation of natural language inference on the Heuristic Analysis for NLI Systems (HANS) dataset (McCoy et al., 2019.)

Python BERT Related Term Project Suggestions

An illustration of a BERT related Term Projects.

- **Question & Answering** (Q&A) system - Build Q&A, select a few pages of text to analyze and provide answers (most relevant sentences from the text) to the posed questions. For example,



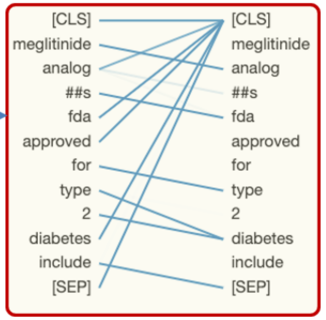
```
question = "Who ruled ancient Macedonia?"

text = """Macedonia was an ancient kingdom on the periphery of Archaic and Classical Greece,
and later the dominant state of Hellenistic Greece. The kingdom was founded and initially ruled
by the Argead dynasty, followed by the Antipatrid and Antigonid dynasties. Home to the ancient
Macedonians, it originated on the northeastern part of the Greek peninsula. Before the 4th
century BC, it was a small kingdom outside of the area dominated by the city-states of Athens,
Sparta and Thebes, and briefly subordinate to Achaemenid Persia."""

# ==================
# Question: Who ruled ancient Macedonia?
# Answer: the Argead dynasty
# ==================
```

- Have a look at https://qa.fastforwardlabs.com/pytorch/hugging%20face/wikipedia/bert/transformers/2020/05/19/Getting_Started_with_QA.html or similar example online. Understand the code and build a Q&A System using any of the BERT models that will provide the answer from the Wikipedia content. For example:

  Question: “Why is the sky blue?”

  Answer pulled from Wikipedia content: “Rayleigh scattering”

Bert Model Awareness of what the next word should be



For any of these BERT projects use visualization of BERT layers: https://github.com/jessevig/bertviz

For visualization of BERT layers, use: https://github.com/jessevig/bertviz

Python BERT Related Term Project Suggestions

Other examples:

- **Question Generation** - Explore the provided “Question Generation” code that generates questions from given text. Integrate this with and Q&A code and check if it returns the right answer.
  - Example: for the following input text
    - "Sachin Ramesh Tendulkar is a former international cricketer from India and a former captain of the Indian national team.”
    - "He is widely regarded as one of the greatest batsmen in the history of cricket.”
    - "He is the highest run scorer of all time in International cricket.”
  - These are the questions that were generated:
    - “What is Sachin Ramesh Tendulkar’s career?”
    - “Where is Sachin Ramesh Tendulkar from?”

- **Sound to Text** - record yourself reading a page of text (TED Talk for example) as wav file and then use the provided “Wav2vec” code to turn it into a text that you can summarize using any NLP summary method. Estimate correctness of the voice to text transcription (how many corrections to the text you needed to make) and illustrate what was the summary of the initial text.

- **Keywords Extraction** - Use provided code to extract keywords from few pages of text (TED Talk for example). Then summarize the text by different methods and extract the keywords again. Compare if the quality of the extracted keywords improved.

- Text sentiment, Text classification, Semantic Similarity, Next sentence prediction…
  - Use visualization of BERT layers: https://github.com/jessevig/bertviz

Preparing and analyzing the data

- If you use data of your choice indicate and document clearly what you have used, especially if the data content changes over time (i.e. stock quotes).

- Document the import the data set into R if necessary, for the project.

- Document the steps for the import into R if necessary, for the project and if any preprocessing had to be done prior or after the import.

- Please keep the naming convention of the R objects if indicated.

- Save these R objects (only if indicated) and include them as files in your submission of the term project.

All the term projects are worth 100 points. The BERT projects will carry extra credit.