CS410 Project Proposal

Xiaolong Yang

Topic Selection:

Recommendation combined Topic Models & Neural Language Models on Wikipedia data.

1.What is the function of the tool?

The tools aimed to build recommendation system for Wikipedia.

Wikipedia is a free online encyclopedia. It is one of the famous Big Text data source and used in many text mining related project, such as Google Knowledge Graph, Wikipedia Search Engine. I hope to build a tool to automatedly recommend relevant Wiki page on user's interested page, such as recommend President Trump on President Obama's page.

2.Who will benefit from such a tool?

Anyone use Wikipedia might benefit from the recommendation tool.

Each Wikipedia page is a topic, you can easily find related information within the topic. Take President Obama as example, you can find his personal details, early life and career, presidential campaigns, presidency and so on in his wiki page. Wikipedia use hyperlink in the page to indicate other topics related to current topic. In the President Obama example, you can find "United States presidential election, 2008" within the wiki page, which is highly related events. However, there are also meaningless topics such as "English", "Seattle", which is the language he knew or where he stayed for a while.

In this case, if I can develop a tool to show merely the highly related topics of user interested, it would save user's time to explore the document.

3.Does this kind of tools already exist? If similar tools exist, how is your tool different from them? Would people care about the difference?

The Google Knowledge Graph had similar function, and was also built based on Wikipedia. Knowledge Graph is a knowledge base used by Google for semantic-search. It uses a

graph database to provide structured and detailed information about a topic and assemble related other topics. If you search for a topic, such as "Barack Obama", it would show his biology, recommend related topic such as "Donald Trump", "Hillary Clinton", his vice president and Children. However, since the Knowledge Graph mostly used Semantic Link Network method to render the knowledge, it used less Text mining techniques. In this case, I hope to build a knowledge graph based on text mining techs.

I would build my recommender with two techs, Topic Models & Neural Language Models. As I mentioned in Project topic selection:

The main difference between these models is the contextual information they use: LSA and topic models use documents as contexts, and Neural language models and distributional semantic models instead use words as contexts. These different contextual representations capture different types of semantic similarity; the document-based models capture semantic relatedness (e.g. "boat" – "water") while the word-based models capture semantic similarity (e.g. "boat" – "ship").

For President Obama example, I hope to find "President" topic using topic models, and "Donald Trump" topic using neural language models.

4.What existing resources can you use?

I find the Wikipedia data source in https://dumps.wikimedia.org/enwiki/20170620/, and would use part of the data for my tools. I might use built-in tools for word2vec in https://www.tensorflow.org/tutorials/word2vec.

5.What techniques/algorithms will you use to develop the tool? (It's fine if you just mention some vague idea.)

I would use Topic Models algorithm like LDA and Neural Language Models like Word2Vec.

6.How will you demonstrate the usefulness of your tool.

If the tool is successful, user can use the tool to easily find highly related topic in Wikipedia to save their time, or find different but also important information that Knowledge Graph cannot tell.

7.A very rough timeline to show when you expect to finish what. (The timeline doesn't have to be accurate.)

Week 7-8 Clean and find a reasonable target to analyze

Week 9-10 Try different Topic Model and Neural Language Model, tune parameters

Week 11-12 Build a website or app to represent the result

Week 13-14 Final report.