Generating Surveys of NLP and AI Topics from Slides using Extractive Summarization Techniques

Project Description
This project aims to employ extractive summarization techniques that will be able to produce comprehensive, coherent summaries of a singular topic using text resources extracted from multiple sets of teaching slides that touch upon the topic. I investigate methods on extracting salient information from condensed, disconnected information to automatically generate a survey of the subject that covers all the key factoids.

This project will make use the AAN database to access the teaching slides focused on NLP and AI, these resources are then annotated for their relevance to certain NLP topics, as well as their prerequisite knowledge topic.

Different extractive summarization techniques will be implemented to compare their results in terms of their scoring of salient paragraphs. There is already a body of existing literature that addresses extractive summarization techniques for a specific genre of resources such as academic papers or news articles. This project will build on this existing literature, and attempt to adapt these techniques to apply to text parsed from pdf slides. Given time, the output will also be compared against the performance of these techniques on a single genre compared to a mixture of genres (including web tutorials or short pdf summaries), and include images by parsing their captions for relevance. Different scoring criteria will also be experimented with to try and generate the best output. This system will be developed using Python, with the goal of ultimately integrating it with the AAN interface. The project will ultimately culminate in a short paper detailing the output, an analysis of the output, and the process of creation.

To measure the performance of the deliverables for the project, the output will be compared against existing wikipedia articles in terms of the percentage of key factoids covered and percentage of non-relevant information. Wikipedia was chosen as a source mainly because a lot of the existing articles in the realm of NLP are well written and comprehensive, and serve as a human control group in summarization.

Data sets
AAN database
- resources
- parsed text files for pdfs can be found here:
  /home/lily/af726/projects/AAN_corpus/data/cards
- topics_annotations

Timeline
1/29 - 2/22 Phase 1: Annotation of resources and familiarization with previous literature.
This phase will consist of the preparation of the data, where I will help with annotation along with the rest of the Survey Generation team. During this time, I will also familiarize myself with summarization in theory and application, in particular extractive summarization and multi-document summarization.

2/23 - 4/7 Phase 2: Implementation of different summarization techniques.
This phase will consist of implementations of a few of the methods of summarization studied in the previous literature, applied to our dataset of annotated resources.

3/30 - 4/28 Phase 3: Analysis of comparisons and output

This phase analyses the output created by the different models and compares them, and visualizes some of the differences. A qualitative analysis of the summaries will also be made.

If the tasks in this timeline is completed ahead of time, additional models can be added to the analyses, or more tweaking can be done to the existing implementations to get better outputs.