Homework 2

Due: October 1, 2018 11:59pm EST

Identifying technology trends is of core importance to venture capitalists, companies and individuals who may invest money or time to pursue the hottest areas. Using historic data, the goals are to characterize either an increase or decrease in certain areas over a span of time, and use that information to predict the next areas before everyone becomes aware of the trend. Economists and financial traders routinely develop methods to achieve this goal using numeric data, but that's a different problem.

Mining published literature for trend detection is not a new area, but it is far from being adequately solved. There are a number of papers that describe case studies for a given area, but none offer a definitive approach; most focus on only a niche area. The two main approaches to the problem are using word concepts and citation networks. The word concept approach aims to characterize a subfield by its component terms automatically and then look for patterns over time. Google Trends offers a plot of word frequency over time, but subfields tend to be more complex in that "convolutional neural network" has synonyms or abbreviations (CNN) that can be ambiguous. Furthermore, as areas mature, the concepts may refine into distinct groups and associate with specific sets of terms. The citation network looks for patterns in which authors are referenced to characterize concepts. These can be used to separate different areas based on which paper is cited, but also tend to be fairly noisy.

This homework will give you and a *required* partner a chance to develop your text mining skills to computationally find the top 10 upward or downward trending areas within the context of 30 years of the Neural Information Processing Systems (NIPS) proceedings for their annual conference.

Data set:

The official data set is the NIPS Proceedings available at https://papers.nips.cc/. However, this will take a long time to download and hammer their server so we will would like to provide you with alternatives. There is a version of the dataset here: https://www.kaggle.com/benhamner/nips-papers. You will need a Kaggle login in order to download it. Since I would prefer everyone spend more time on the analysis and less time on the cleaning, I am working to put out a slightly cleaner version of the official data set shortly that I will post.

Partners:

HW2 is a partnered homework so work should be completed with 1 partner. To help everyone find a partner, we ask you to sign up by putting your partner's first name next to yours and vice versa using this shared spreadsheet: https://docs.google.com/spreadsheets/d/1oz0pNYx8X2WptwiLsD9zMUtsCVZiEUTFXZ5DEnPaewk/edit?usp=sharing. This will give everyone immediate feedback on who doesn't have a partner.

To select a partner, the self-intros on piazza are a good place to start. Please use the Canvas email to contact them since we respect your privacy and don't want to post everyone's email.

Suggestions on strategies:

You are welcome to pursue any approach. If you find applicable methods online, feel free to use them and be sure to cite the results. I would recommend starting with the text mining pipeline described in lecture and section to clean the documents and identify single- and perhaps multi-word terms. In this case, the first pass might be to perform simple counting as a baseline over time and work for a standard approach to plot trends taking the normalization into account. In the next pass, you might expand from the isolated word terms to synonyms to larger concept subfields that may cluster together. The citations or co-related words can be helpful for this. Further refinements might be to include only certain sections of the documents or try weighting schemes.

Grading philosophy:

We will grade based on 1) your success in the project so label your final result, and 2) your exploration of different ideas. Please document your success, but also document your rationale and failed approaches. We want to know which hypotheses you pursued and how they panned out. With these kinds of homework, we expect both partners to work together and contribute equally to a greater result than either could do alone given the time constraints. We will post a form to assess your partner's contribution relative to yours.

What to Submit:

Please submit your python notebook and associated pdf of that notebook. In a separate document, please also submit a brief description (1-2 paragraphs each) as a separate document to address the following:

1. How have you defined a trend? How can you separate it from background noise and/or spurious relationships?
2. What are the main techniques you have used and how have you tailored them for this problem?
3. What was your strategy for finding multi-word phrases versus single words?
4. What approach(es) did you use to separate one subfield from others?
5. What parts of the document did you use and why?
6. How did you normalize the results against the growth of the conference, lengths of documents, etc.?
7. We know that you can look back and find trends but how would you find the next trend with your method? Be specific.
8. Plot of the final top 10 normalized trends as a function of time.

To assist the grading within the notebook:

- Label your final approach within the file for grading purposes.
- Flag the distinct approaches with a header describing your strategy and corresponding results. It makes it much easier to follow your rationale with headers and descriptions than trying guess using the code alone.

We hope that you find this to be an interesting problem.