1. How have you defined a trend? How can you separate it from background noise and/or spurious relationships?

We defined a trend as a word or multi-word phrase that shows increasing relative frequency over time. We use linear regression to calculate the rate of change of relative frequency (slope) and choose those words with the highest slope. By doing this, we focus on words with the biggest increase in relative frequency, which should be separated from background noise and/or spurious relationships.

1. What are the main techniques you have used and how have you tailored them for this problem?

The first main technique we have used is linear regression with relative frequency of words as a function of time. Specifically we used CountVectorizer for preprocessing and frequencies counting. We extracted both unigrams and bigrams, and analyzed both abstracts and full texts. We then normalized the frequency of each word to the length of the article as the relative frequency and calculated the average relative frequency of each word for each year. We conducted a linear regression with average relative frequency of words as a function of time and selected words with highest positive slope. These words show increasing relative frequency over time and thus represent a trend.

1. What was your strategy for finding multi-word phrases versus single words?    
   To start with, we simply set ngram = 2 in CountVectorizer to find bigrams. Then we manually curated bigrams with increasing relative frequency to select bigrams that represent real phrases.
2. What approach(es) did you use to separate one subfield from others?

LDA? Please add more, Noah.

1. What parts of the document did you use and why?

We used both abstracts and full texts and integrate hits from both sources. However, we give higher weights to words identified from the abstract. An abstract summarizes the main idea of an article in a concise and accurate manner, and therefore provides much higher fraction of key words that may represent a trend. On the other hand, full texts include more information, which sometimes dilute out the signal of key words that represent a trend.

1. How did you normalize the results against the growth of the conference, lengths of documents, etc.?

For the linear regression approach, we normalized the frequency of ngrams to the length of articles and take the average of relative frequency of ngrams for each year. We think normalization to the lengths of articles and averaging relative frequencies of words for each year should minimize confounding effects from the growth of the conference and lengths of documents.

1. We know that you can look back and find trends but how would you find the next trend with your method? Be specific.

The words that represent trends all exhibit increasing relative frequency over time, so we should be able to use time series to forecast their future relative frequency. If they still show increasing relative frequency, they probably also represent trends in the (near) future.

1. Plot of the final top 10 normalized trends as a function of time.

Please compare the results from the linear regression with your LDA results and choose those that make sense to you as final top 10