Name:
Behrouz Ebrahimi
Sharjil Khan

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

We defined a trend to be either an “upward trend” or a “downward trend”.

Upward Trend**:** Any term or idea that has seen an increase in the number of times it was mentioned (normalised) over the last 3 consecutive years.

Downward Trend**:** Any term or idea that has seen a decrease in the number of times it was mentioned (normalised) over the last 3 consecutive years.

In addition to the above constraints we also looked at the terms that have featured in the top 20 most important terms in at least one of those 3 years. This allowed us to ignore terms that had low overall importance.

We also filtered out terms that are unrelated to Machine Learning and AI.

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

We first used clustering and LDA to try to group terms into subfields and assigned all the other terms to their relevant subfields. We also tried to group the individual articles into groups to see which fields each article represent. Even though we were able to gain some good insights using the above techniques, we believe the groupings were too ambiguous in most cases and we were probably grouping terms into groups that actually belong to different fields.

Because of the above concern, we eventually decided to look at multi word terms as fields and plot their trends without trying to group them in any way. The results showed some good trends and we believe it was less subjective because in this method we were only counting the words as they were seen in the papers for each year.

We then found the upward trending and downward trending topics plotted them against years.

3. What was your strategy for finding multi-word phrases versus single words?
We used single words and multi words to find topics using LDA.

Trying out single word terms Vs two word terms revealed that the two word terms where much better at revealing machine learning related ideas and concepts. So, we decided to focus on

only two word terms in our final TFIDF analysis because they were much better at capturing key ideas and concepts.

4. What approach(es) did you use to separate one subfield from others?
In our initial approach we used Latent Dirichlet Allocation to separate topics and then looked at the words and articles in each topic to give them the most relevant field name.

From then onwards we gave each article in the corpus a weight for each topic based on the probabilities of them being assigned to that topic. We used the mean of those weights to plot the subfields for each year.

In the second part, we used two word terms to depict topics and used their raw tf idf scores to plot them. We realised that two word terms were quite effective in encapsulating fields and areas in the field of machine learning and AI.

5. What parts of the document did you use and why?
We used the title and the full text for our analysis. Because the full text already contains the abstract in it, we did not use the abstract because that would result in double counting the words in the abstract.

We also applied a 3x multiplier to the terms found in the title because we believe if a certain term appears in the title it should be given more weight then if it appears in the text of the paper.

6. How did you normalize the results against the growth of the conference, lengths of documents,
Etc.?
We know TFIDF penalises the scores for a term if it is too common across the entire corpus. Because of this, we did not group our corpus by year while measuring the TFIDF scores of our terms. By using the entire corpus from 1976 to 2017 we made sure that if a term is very common for a particular year, it would not get penalised because we want to be able to capture the fact that a term was very popular for a specific year.

After calculating the TFIDF scores for each term, we took the mean of the TFIDF scores for each year. This took care of the fact that the number of documents per year was increasing for the later years as a result of the growth of the conference.

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

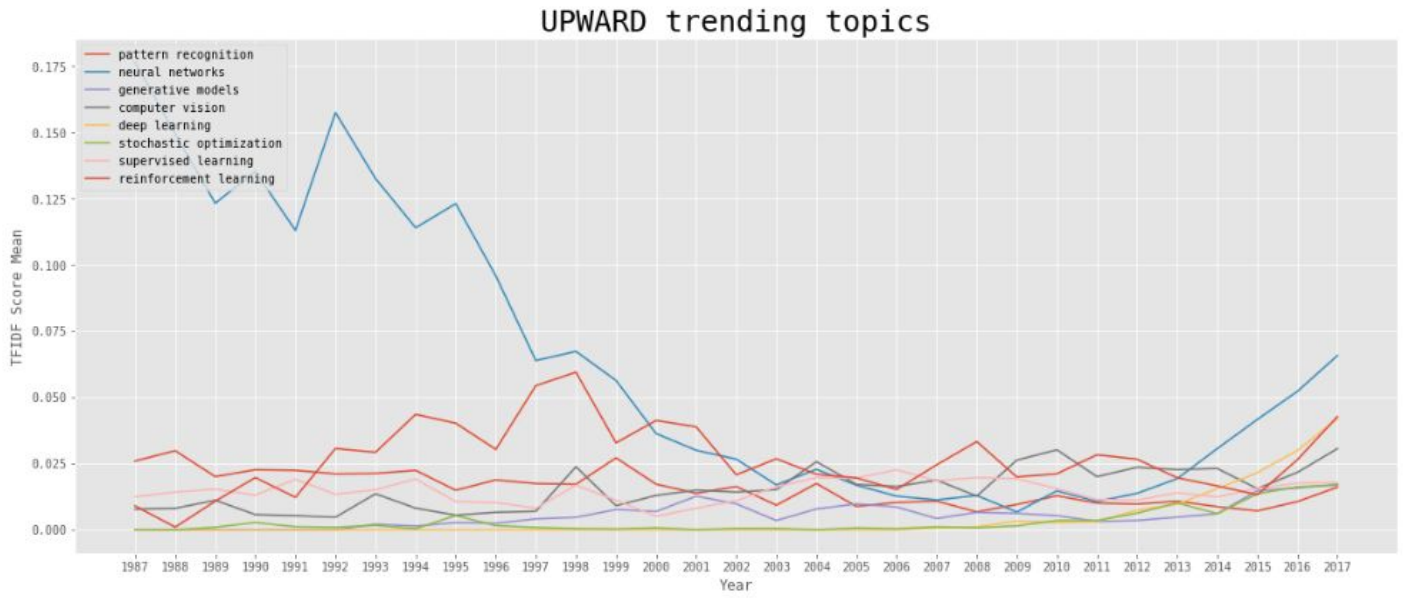
We would start with a table of average TFIDF scores for each year with the columns depicting years and each row containing individual terms/fields.

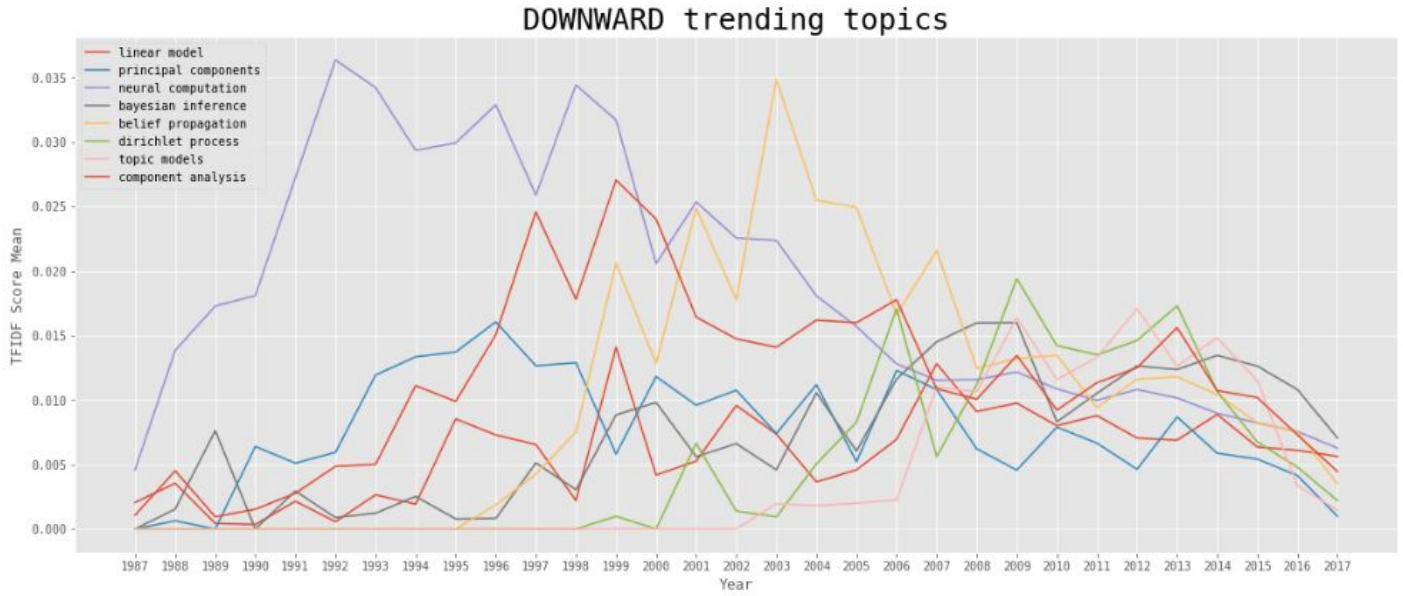
Then we would convert the last 3 years that we used to find a trend and convert it to a binary value depicting a trending or a non-trending topic.

We would then use this data to train a neural network. The input layer will be the values for all the years prior to the 3 years that we used to figure out if its a trending topic. The output layer will be the binary value “trending” or “not trending”.

We can then use this neural network to predict a trending topic by using the tfidf average for all the years prior to the point from where we want to know if it will be a trending topic or not.

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


UPWARD trending topics
pattern recognition
neural networks
generative models
computer vision
deep learning
stochastic optimization
supervised learning
reinforcement learning
TFIDF Score Mean
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
1987 1988 1989 1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017
Year


DOWNWARD trending topics
linear model
principal components
neural computation
bayesian inference
belief propagation
dirichlet process
topic models
component analysis
TFIDF Score Mean
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
1987 1988 1989 1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017
Year