

# Dynamic Academic Hot Topic Discovery

Chen Liu 2011010539  
Yang Liu 2011012162

## Our webservice: dminer.duapp.com

Our webservice offers a convenient way to detect hot academic topics and their relevance over time. We provide two APIs for users to query, one is to present related topics in a quantity way of the input over time, the other is to draw a chart line showing the dynamic relevance of two academic topics users have input.

Our webservice is available at dminer.duapp.com, which is powered by baidu application engine. We use python flask framework to implement our webservice. In the backstage, we take advantage of several machine learning algorithms to train massive information and store the result model in a database customized for python flask framework. In the website, we use javascript and several library like ichart to communicate with backstage and make it user-friendly.

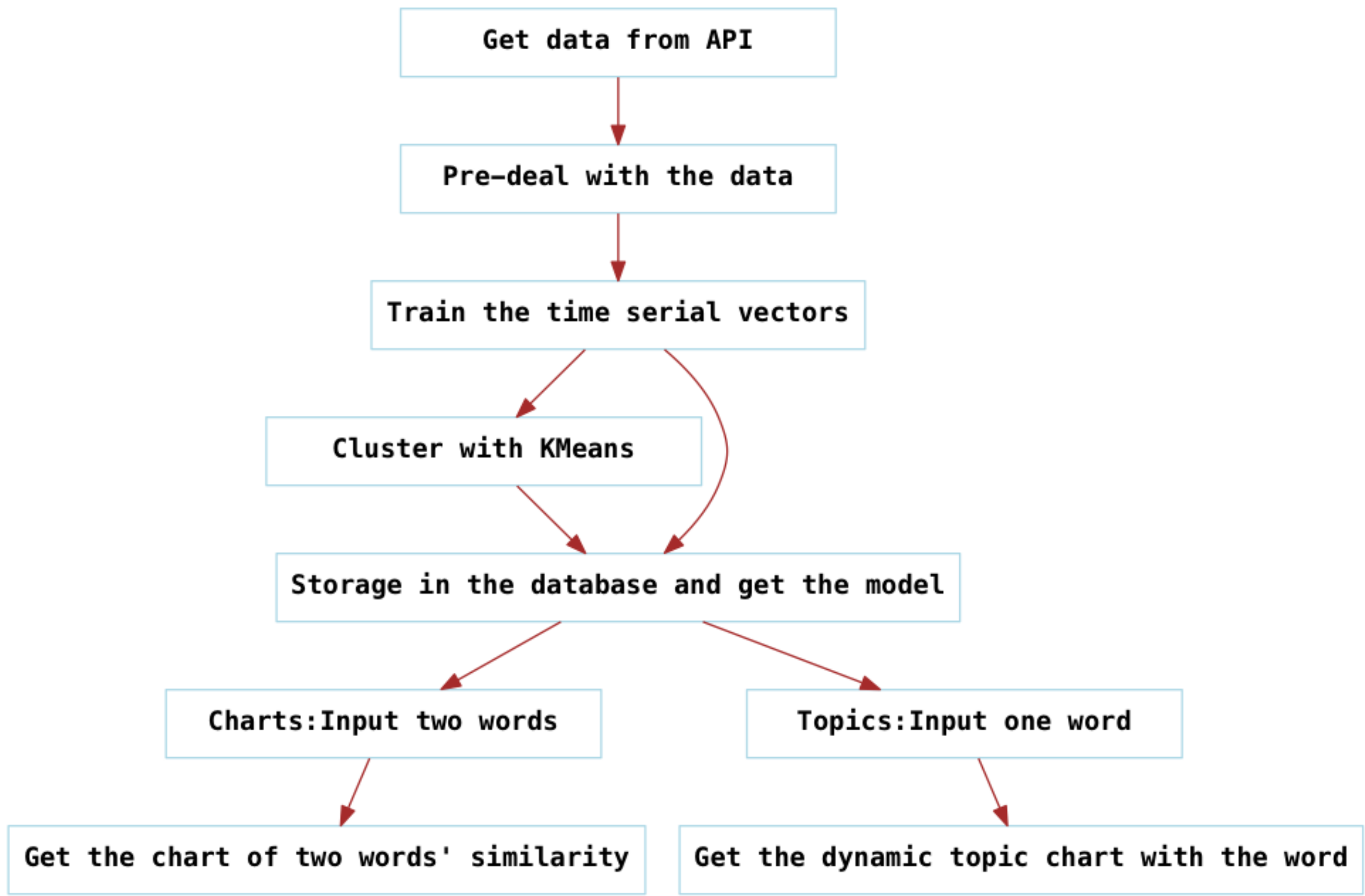
We use the model of Word2vec to describe the word and represent the words in the paper or the passage by the vectors in the real vector space and train the data using the time serial data.

After training, we get the vectors of word among the time and do a cluster of the words. Getting the classes of the paper through different years, we can do the discovery of the hot topic changing through the time.

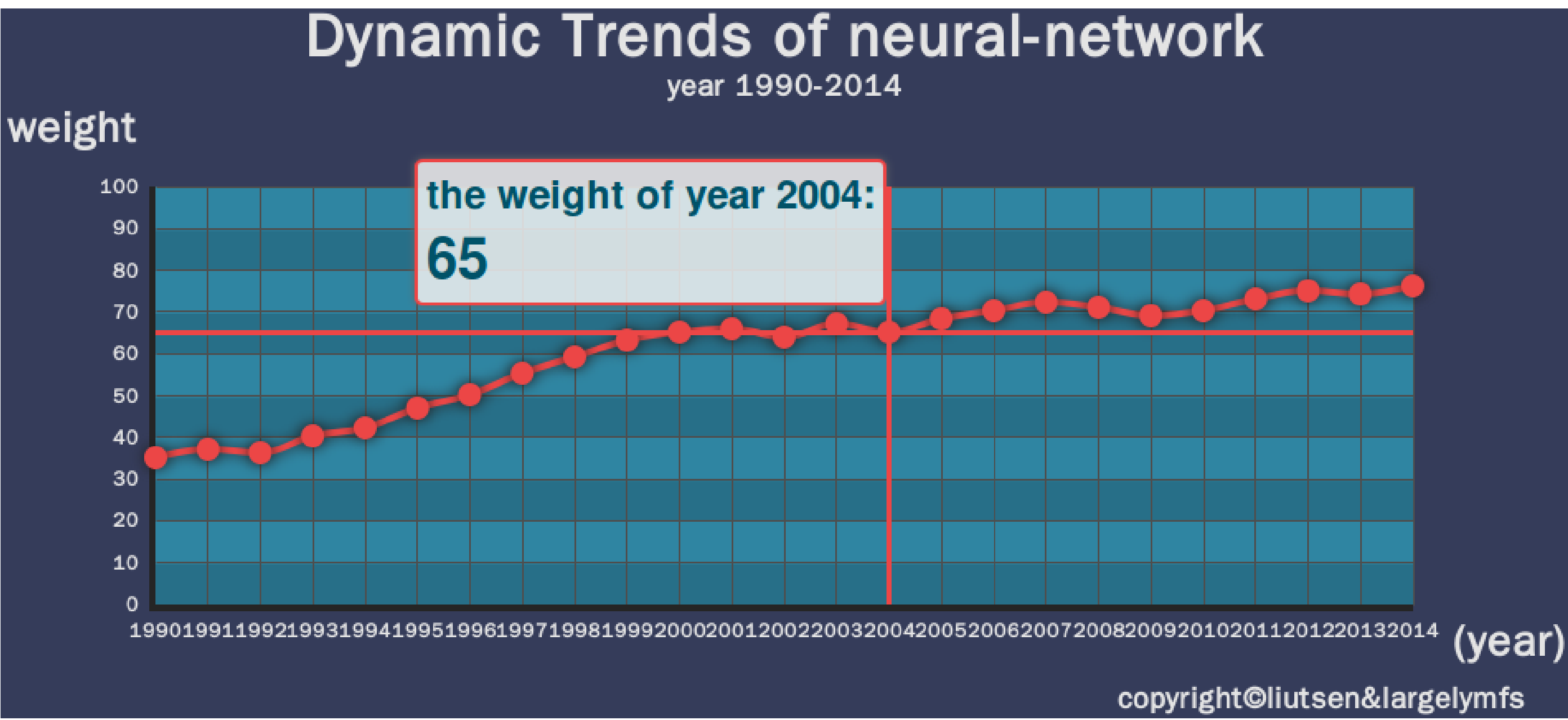
We storage the word vectors and the classes in the database for the user. Using the two API available, we can get the similarity changing of two words and the hot topic about the key word.

For demo, we try to use the NIPS(Advances in Neural Information Processing Systems) paper at <http://books.nips.cc/papers/txt>, and the data of Science Magezine:<http://www.sciencemag.org/cgi/collection/physics>

The demo system’s process is as the following graph:

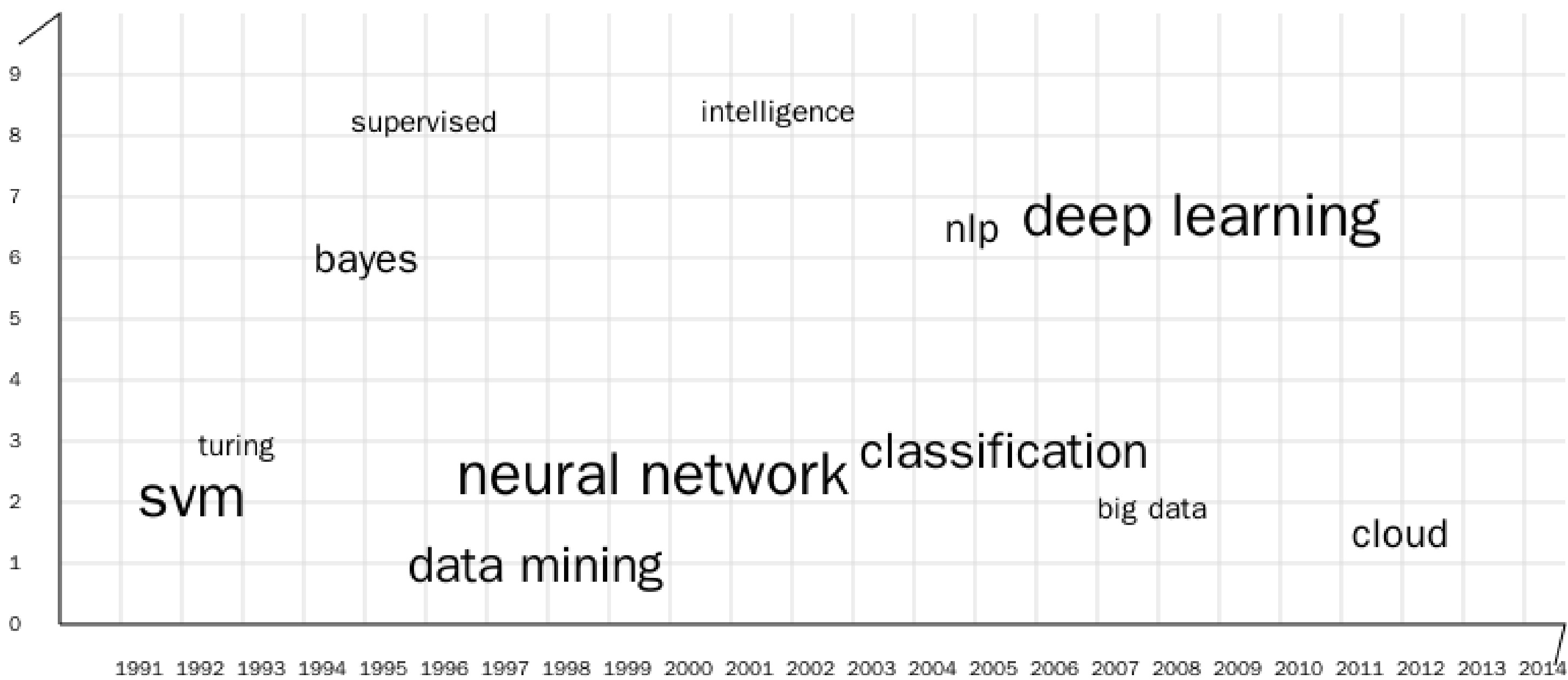


The procedures of the demo



The line chart showing the relevance of two topics over a period of time

As is shown above, the user types in two words and the engine fetches the coresponding vector from the database, the backstage calculate the cosine of the two vector in different period of time and normalize them into a scale of 0 to 100. The backstage send the information to the website in the form of string and the website parse it to render the line chart framework.



The graph showing the related topic over time

Another api is to show the relative topics of the input keyword over time. One of possible result is the picture above (input='machine learning'). The backstage calculates the cosine distances between the input word and the word of the same clusters. Then, we sort the words of same cluster by distance and find out their booming time. The website receives the information and renders the canvas framework so that the words are sorted by time and the closer words appear bigger.

## Reference:

- 1.ichartjs library: <https://github.com/wanghetommy/ichartjs>
- 2.modernizr library: <http://modernizr.com/>
- 3. NIPS(Advances in Neural Information Processing Systems) <http://books.nips.cc/papers/txt>
- 4. Science Magezine:<http://www.sciencemag.org/cgi/collection/physics>
- 5.Word2vec : <http://code.google.com/p/word2vec/>