一个自然语言处理的小例子，选择最相似的前n门课

# **初始课程**



总共379个课程，每行包括3部分内容：课程名\t课程简介\t课程详情, 已经清除了其中的html tag, 下面所示的例子仅仅是其中的课程名：

*Writing II: Rhetorical Composing  
Genetics and Society: A Course for Educators  
General Game Playing  
Genes and the Human Condition (From Behavior to Biotechnology)  
A Brief History of Humankind  
New Models of Business in Society  
Analyse Numérique pour Ingénieurs  
Evolution: A Course for Educators  
Coding the Matrix: Linear Algebra through Computer Science Applications  
The Dynamic Earth: A Course for Educators  
...*

# **预处理**

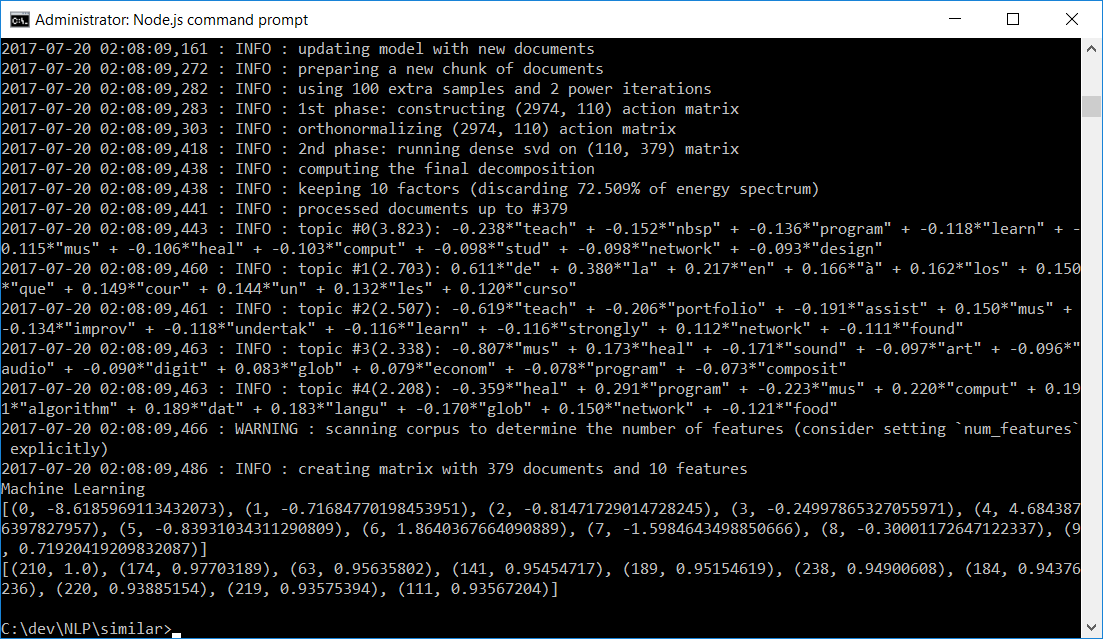
* 文档的单词小写化
* 标点符号和单词是没有分离的 （tokenize）
* 去停用词 （去掉分词，介词 。。。）
* 去掉标点符号
* 英文单词词干化
* 去掉在整个语料库中出现次数为1的低频词

# BOW模型，建立词袋

# 计算TFIDF

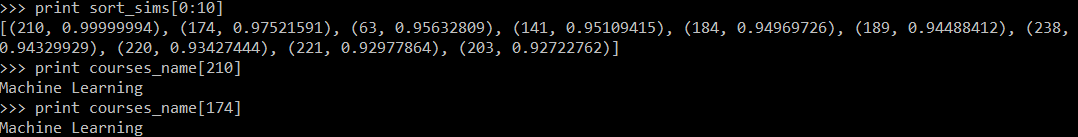
# 训练LSI模型

# 寻找和Machine Learning课最相似的前十门课



# 最相似的前十门课输出是

[(210, 1.0), (174, 0.97703189), (63, 0.95635802), (141, 0.95454717), (189, 0.95154619), (238, 0.94900608), (184, 0.94376236), (220, 0.93885154), (219, 0.93575394), (111, 0.93567204)]



第一门课程是它自己:  
>>> print courses\_name[210]  
Machine Learning

第二门课是Coursera上另一位大牛Pedro Domingos[机器学习公开课](http://coursegraph.com/machine-learning-coursera-machlearning-university-of-washington)  
>>> print courses\_name[174]  
Machine Learning

第三门课  
>>> print courses\_name[63]

Human-Computer Interaction

第四门课>>> print courses\_name[141]

Computational Photography

ToDO

* 中文
* Google Tensorflow.需要时间熟悉百度的AI框架及其效果
* 预处理步骤很关键，数据质量不佳会导致学习效果的下降

