《计算社会科学》授课大纲

预习安排

- Anaconda、Jupyter Notebook 等的安装
- 微积分、概率论、线性代数基础知识
- request、beautifulsoup 库的使用（也许 re 等其它爬虫相关的库）
- 具体日程待定

Day1：计算社会科学简介

- Lead in：
  - 案例：《千面英雄》→所有英雄史诗的叙事结构；公元纪年以来各个时间段中在历史上留名的人物的职业。
  - 费曼学习法（为点名埋下伏笔）→自我介绍、时间安排
- 一、计算中心的逻辑
  - 在 21 世纪，约 47%的工作将有大于 70%的概率被自动化，而科学、教育等领域则可能性较小。
  - 大问题：人们被局限在自己日常生活的轨迹中，缺乏对整个世界的感知，在这样的情况下，他们觉得自己的生活是一连串的陷阱，越是以旁观者的视角去看这个世界，他们就越是觉得自己被困住了。《社会学的想象力》
  - 在计算中心的逻辑下，我们应当让局限于当地的人获得全球的知识，因为理论在被提出之后就成了一个黑箱，我们需要围绕着计算使这成为可能。
  - 计算社会科学领域划时代的论文：*Lazer, D. et al. 2009. Computational Social Science. Science. 6 February 2009: Vol. 323, no. 5915, pp. 721-723.*

“简而言之，计算社会科学是一个新兴的领域，它以前所未有的广度、深度和规模利用、收集和分析数据。”

“数字痕迹可以被汇编成个人和群体行为的全景，有可能改变我们对我们的生活、组织和社会的理解。”

但是问题在于：1、出现的晚发展的慢；2、大公司与政府垄断数据；3、进而导致研究可能难以复现、检验。4、目前的研究范式不适应。5、研究伦理与隐私问题

  - 计算社会科学的领域
    - （大）数据与自动数据提取
    - 社会网络
    - 社会复杂性
    - 仿真模拟
- 二、数据科学的编程工具

Jupyter Notebook、机器学习及相关的书籍推荐

- 三、计算社会科学的研究方法

由相关关系到因果关系：相关→干预→反事实分析

案例介绍；伦理问题（剑桥分析事件）

Day1 作业

- 1、使用 mobirise 建立个人网站
  下载 mobirise 软件 https://mobirise.com/
  制作个人网页上传到 Github
  例如 https://chengjunwang.com
- 2、下载并分析 ows-raw.txt 数据，统计每一天有多少人多少 tweets。

参考资料

1. *Cioffi-Revilla, Claudio (2014). Introduction to Computational Social Science. Springer-Verlag, London.*
2. *Matthew J. Salganik (2017) Bit by Bit: Social Research in the Digital Age. Princeton University Press*
3. *张伦, 王成军, 许小可 2018 《计算传播学导论》北京师范大学出版社*
4. *VanderPlas, J. (2016). Python data science handbook: Essential tools for working with data. " O'Reilly Media, Inc.".*
5. https://github.com/SocratesAcademy/machine-learning（机器学习）

参考文献

1. Yu, A. Z., et al. (2016). Pantheon 1.0, a manually verified dataset of globally famous biographies. Scientific Data 2:150075. doi: 10.1038/sdata.2015.75
2. Frey, C. B., & Osborne, M. A. (2017). The future of employment: How susceptible are jobs to computerisation?. Technological forecasting and social change, 114, 254-280.
3. Lazer, D. et al. 2009. Computational Social Science. Science. 6 February 2009: Vol. 323, no. 5915, pp. 721-723.
4. D. Watts, A twenty-first century science. Nature 445, 489 (2007).
5. Conte, R. 2012. Manifesto of Computational Social Science. The European Physical Journal Special Topics. November 2012: Vol. 214, Issue 1, pp. 325-346.
6. Cioffi-Revilla, Claudio (2014). Introduction to Computational Social Science. Springer-Verlag, London.
7. Gary King (2011) Ensuring the Data-Rich Future of the Social Sciences. Science.
8. Watts, D. J. (2014). Common sense and sociological explanations. American Journal of Sociology, 120(2), 313-351.
9. Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D., Marlow, C., Settle, J. E., & Fowler, J. H. (2012). A 61-million-person experiment in social influence and political mobilization. Nature, 489(7415), 295-298.
10. Salganik, M. J., Dodds, P. S., & Watts, D. J. (2006). Experimental study of inequality and unpredictability in an artificial cultural market. science, 311(5762), 854-856.
11. Culotta, A. (2010, July). Towards detecting influenza epidemics by analyzing Twitter messages. In Proceedings of the first workshop on social media analytics (pp. 115-122).
12. Blumenstock, J., Cadamuro, G., & On, R. (2015). Predicting poverty and wealth from mobile phone metadata. Science, 350(6264), 1073-1076.
13. Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. Science, 359(6380), 1146-1151.

Day2 大数据

- 大数据概要
  - 大数据定义和用途

（观察数据 observational data、罕见事件的研究、对异质性的研究、检测微小差异、实现除研究之外的其他功能、3Vs、5Ws）

【注意事项：减少随机误差，增加系统性误差，需关注数据的创建过程】

  - 大数据来源

（企业：信息设备、政府：记录）

- 大数据特征

（海量性、持续性、不反应性、不完整、难获取、不具代表性、漂移、算法干扰、脏数据、敏感性）

    - 功能：unexpected events, real-time estimates、人们不会改变行为
    - 弊端：无法避免社会期许偏差、用户行为受到企业算法影响、找不到需要的信息
    - 解决方案：收集所需、用户属性推断、结合多个数据源（记录链接）
- 大数据如何做研究？（好的研究不取决于计算的复杂性或数据本身，而是取决于问题）
    - 计算、预测、近似实验（自然实验、匹配实验）

参考文献：

1. Jean-Baptiste Michel et al. (2011) Quantitative Analysis of Culture Using Millions of Digitized Books.Science 331, 176
2. Raj Chetty, et al. (2017) The fading American dream: Trends in absolute income mobility since 1940. Science. 356(6336):398-406. https://opportunityinsights.org/

Day3 数字时代的调查与实验

- 调查与访谈
    - 调查有关要素：大量参与者、高度结构化的问卷调查、统计学以样本推断总体
        - 数字时代：数字时代如何影响了调查研究？

（快速且低成本地收集信息、大数据放大调查的价值、Wiki 调查……）

        - 调查研究用途（了解潜在混杂因素、测量主观感受、收集非在线行为与互动信息）
        - 调查误差（TE=RE+ME）
            - 覆盖误差、抽样误差（非概率抽样、概率抽样、需事后分层 post-stratification 做加权平均来估计总体、运用组内同质反应密度假设）、无回答误差（应答比例过低）、测量误差（从受访者的回答中推断想法和行为、需防止社会期许偏差和采访者影响、提问方式效应、需阅读原始问卷、使用 EMA 法）
    - 深度访谈有关要素：少量参与者、半结构化的对话、丰富的定性的描述
- 实验
    - 因果推断
        - 因果推断的本质是什么？（反事实、潜在结果框架）
    - 什么是实验研究？包含哪些步骤？有何优势或劣势？（随机化、实验的两个维度、机制研究、实验技术、实验伦理）
    - 实验类型（实验室实验、实地实验、自然实验、工具变量）

参考文献：

1. Schuman and Presser (1996) Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording, and Context. Thousand Oaks, CA: SAGE.
2. Sugie, Naomi F. 2014. “Finding Work: A Smartphone Study of Job Searching, Social Contacts, and Wellbeing After Prison.” Ph.D. Thesis, Princeton University. https://dataspace.princeton.edu/jspui/handle/88435/dsp011544br32k
3. Salganik, Matthew J., and Karen E. C. Levy. 2015. “Wiki Surveys: Open and Quantifiable Social Data Collection.” PLoS ONE 10 (5):e0123483. https://doi.org/10.1371/journal.pone.0123483
4. Goel, Sharad, Winter Mason, and Duncan J. Watts. 2010. “Real and Perceived Attitude Agreement in Social Networks.” Journal of Personality and Social Psychology 99 (4):611–21.

https://doi.org/10.1037/a0020697
5. Angrist, Joshua D., and Jörn-Steffen Pischke. 2009. Mostly Harmless Econometrics: An Empiricist's Companion. Princeton, NJ: Princeton University Press.
6. Imbens, Guido W., and Donald B. Rubin. 2015. Causal Inference in Statistics, Social, and Biomedical Sciences. Cambridge: Cambridge University Press.
7. Imbens, Guido W., and Paul R. Rosenbaum. 2005. "Robust, Accurate Confidence Intervals with a Weak Instrument: Quarter of Birth and Education." Journal of the Royal Statistical Society: Series A (Statistics in Society) 168 (1):109–26. https://doi.org/10.1111/j.1467-985X.2004.00339.x.
8. Murray, Michael P. 2006. "Avoiding Invalid Instruments and Coping with Weak Instruments." Journal of Economic Perspectives 20 (4):111–32. http://www.jstor.org/stable/30033686.
9. Sekhon, Jasjeet S., and Rocío Titiunik. 2012. "When Natural Experiments Are Neither Natural nor Experiments." American Political Science Review 106 (1):35–57. https://doi.org/10.1017/S0003055411000542.
10. Aronow, Peter M., and Allison Carnegie. 2013. "Beyond LATE: Estimation of the Average Treatment Effect with an Instrumental Variable." Political Analysis 21 (4):492–506. https://doi.org/10.1093/pan/mpt013.
11. Dunning, Thad. 2012. Natural Experiments in the Social Sciences: A Design-Based Approach. Cambridge: Cambridge University Press.
12. Gerber and Green. 2012. Field Experiments: Design, Analysis, and Interpretation. New York: W. W. Norton.
13. Coviello, Lorenzo, Yunkyu Sohn, Adam D. I. Kramer, Cameron Marlow, Massimo Franceschetti, Nicholas A. Christakis, and James H. Fowler. 2014. "Detecting Emotional Contagion in Massive Social Networks." PLoS ONE 9 (3):e90315.

Day4：休息

Day 5. 网络科学

- **学习目标：**
  - 理解关系网络结构（network structure）如何影响复杂系统
  - 理解网络科学工具如何能够量化甚至预测失败
  - **案例介绍：**萨达姆、本拉登、2003 年北美大停电
- **一、网络科学简介**
  - （一）复杂系统 （例如，社会，市场，大脑，细胞）
  - （二）网络研究的历史
  - （三）网络科学的影响
    - 健康：基因研究和疾病治疗
    - 安全：对抗恐怖主义
    - 疫情：预测与阻止病毒传播
    - 大脑研究：绘制神经网络结构
  - （四）网络与图像
    - **无向和有向网络**

定义
无向网络的连接没有指向性，例如论文合作、演员合作、蛋白质反应
有向网络的连接有指向性，例如网址、电话、新陈代谢反应

- 度（Degree）

无向网络中，节点度（Node Degree）是指一个节点的链接数量。
有向网络中，我们需要定义入度（in-degree）和出度(out-degree)。某一节点的度为两者的总和。

$$k_3^{in} = 2,\ k_3^{out} = 1,\ k_3 = 3$$

- 平均度（Average Degree）

无向网络中，平均度为

$$< k >= \frac{1}{N}\sum_{i=1}^{N} k_i = \frac{2L}{N}$$

有向网络中，平均度为

$$< k^{in} >= \frac{1}{N}\sum_{i=1}^{N} k_i^{in} =< k^{out} >= \frac{1}{N}\sum_{i=1}^{N} k_i^{out} = \frac{L}{N}$$

- 度分布（Degree distribution）

$$P(k) = \frac{N_k}{N}$$

P(k)指一个任意选中的节点具备自由度 k 的概率

N_k是指具备自由度 k 的节点数量。

- 邻接矩阵 （Adjacency Matrix）

$$A_{ij} = 1 \text{ 如果节点 } i \text{ 和节点 } j \text{ 之间有一条链接}$$

$$A_{ij} = 0 \text{ 如果节点 } i \text{ 和节点 } j \text{ 之间没有链接}$$

无向网络的矩阵是对称的。
有向网络的矩阵不对称。

- **二分网络（Bipartite Networks）**

二分图像（bipartite graph or biograph）中的节点可以被分成两个离散的子集 U 和 V, 每一条链接联结 U 中的一个节点和 V 中的一个节点。U 和 V 是独立的集合。

- **相关概念**
  - 路径（Path）
  - 距离（Distance）
  - 直径（Diameter）
  - 最短路径（Shortest Path）
  - 平均路径（Average Path Length）

- 环（Cycle）
- 桥（Bridge）
- 结构洞
- 强连接有向图（Strongly connected directed graph）
- 弱连接有向图（Weakly connected directed graph）
- 聚集系数（Clustering coefficient）
- 全局聚集系数（Global Clustering coefficient）

- **二、使用 Network X 分析复杂网络**
  - （一）WWW 数据下载
  - （二）描述网络（网络直径、密度、聚集系数、匹配系数）
  - （三）度中心性测量（Degree centrality meansures）
  - （四）度分布
  - （五）网络结构分析（规则网络、ER 随机网络、小世界网络、BA 网络）

- **三、网络科学模型**
  - （一）随机网络模型
  - （二）度分布
  - （三）泊松分布（Poisson Distribution）
  - （四）小世界现象
  - （五）随机网络的直径
  - （六）六度区隔理论
  - （七）WS 模型（1998）
  - （八）BA 模型（1999）
  - （九）连续平均场（Continuum Mean-Field）

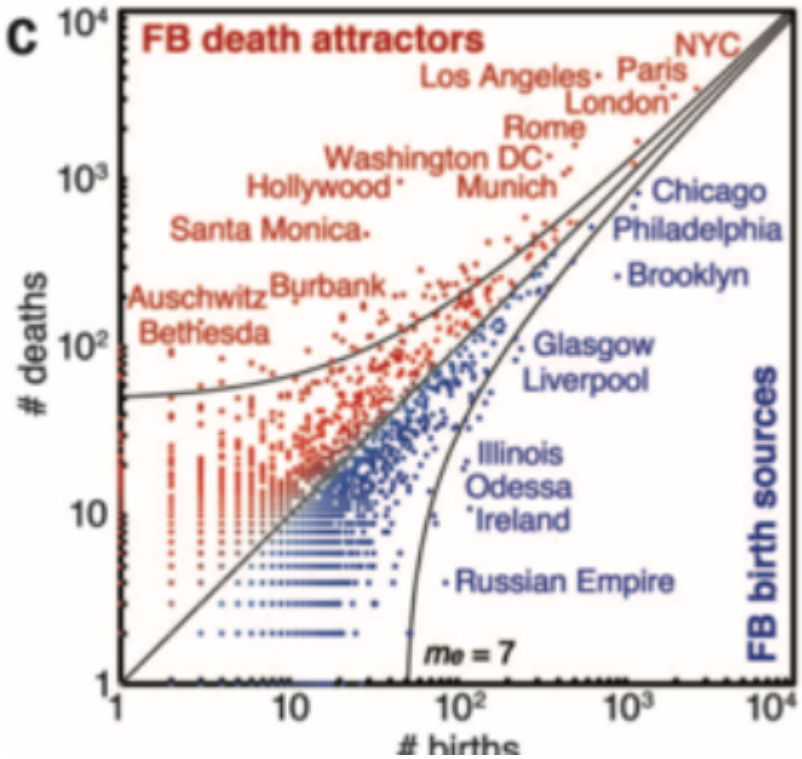
- **四、脸书社交网络的可视化**

Day 5 作业

作业 1

1) 下载 12 万名人出生死亡地理位置数据 Data S1 SchichDataS1_FB.xlsx (Freebase.com)
2) 构建图 D 名人出生死亡的城市网络

3) 绘制度分布曲线
4) 尝试绘制上图 C



作业 2：

1）下载 www 数据
2）构建 networkx 的网络对象 g（提示：有向网络）
3）将 www 数据添加到 g 当中
4）计算网络中的节点数量和链接数量
5）计算 www 网络的网络密度

作业 3：

1) 阅读 Barabasi (1999) Diameter of the world wide web. Nature. 401
2) 绘制 www 网络的出度分布、入度分布
3) 使用 BA 模型生成节点数为 N，幂指数为 γ 的网络
4) 计算平均路径长度 d 与节点数量的关系

网络科学研究论文

1. WATTSDJ,STROGATZSH. Collective dynamics of 'small-world' networks. Nature, 1998, 393(6684): 440–442.
2. BARABÁSI A-L, ALBERT R. Emergence of scaling in random networks. Science, 1999, 286(5439): 509-512.
3. Fan Chung & Linyuan Lu, The average distance in random graphs with given expected degree,. PNAS, 19, 15879-15882 (2002).
4. A.-L.Barabási,R.Albert,H.Jeong Mean-field theory for scale-free random networks. Physica A 272, 173–187 (1999).
5. Liu Y Y, Slotine J J, Barabási A L. Nature, 2011, 473(7346): 167-173.

阅读材料

1. Barabasi 2016 Network Science. Cambridge.
2. 汪小帆、李翔、陈关荣 2012 《网络科学导论》. 高等教育出版社
3. 梅拉妮·米歇尔 2011 《复杂》,湖南科学技术出版社
4. 菲利普-鲍尔 2004 《预知社会：群体行为的内在法则》，当代中国出版社
5. 巴拉巴西 2007 《链接：网络新科学》 湖南科技出版社

参考文献

1. Barabasi 2016 Network Science. Cambridge
2. Barabasi (1999) Emergence of scaling in random networks.Science-509-12.pdf
3. Barabasi (1999) Mean-field theory for scale-free random networks. PA.pdf
4. Albert & Barabasi (2002) Statistical mechanics of complex networks. RMP.pdf
5. Principle of Locality I: Hacking the Continuum Mean-Field Technique in Network Modeling http://www.jianshu.com/p/97f674267d3e

Day 6. 机器学习简介

- **学习目标：**
  - 基于 python 进行机器学习的基本逻辑（需要学员提前安装 anaconda）
  - Scikit-Learn
  - 机器学习模型的参数和模型验证、特征提取。
- **一、什么是机器学习**
  - **（一）监督式学习 (supervised learning)**
    - 目标变量或结果变量（或因变量）y
    - 特征（自变量）X
    - 生成一个将输入值映射到期望输出值的函数 y=f(X)

        - 训练过程会一直持续，直到模型在训练数据上获得期望的准确度。
        - 监督式学习的例子有：回归、决策树、随机森林、K–近邻算法、逻辑回归等。
    - **（二）非监督式学习 (unsupervised learning)**
        - 没有任何目标变量或结果变量要预测或估计。
        - 这个算法用在不同的组内聚类分析。
        - 非监督式学习的例子有：关联算法和 K Means 算法。
    - **（三）强化学习（reinforcement learning）**
        - 训练机器进行决策，例如 Alpha Go。
        - 机器被放在一个能让它通过反复试错来训练自己的环境中。
        - 从过去的经验中进行学习，并利用学到的知识作出判断。
        - 例子：马尔可夫决策过程
    - **（四）泰坦尼克号数据分析**
        - 数据清理
        - 数据描述
        - 数据可视化
        - 基于 Sklearn 使用决策树
        - 基于 Sklearn 使用随机森林

- **二、Scikit-Learn 介绍**
    - （一）Scikit-Learn 中的数据呈现方式
        - 表格、特征矩阵（Features matrix）、目标数组（Target array）
    - （二）Scikit-Learn's Estimator API
        - 监督式学习案例一：简单线性回归
        - 监督式学习案例二：Iris classification
        - 非监督式学习案例一： Iris dimensionality reduction
        - 非监督式学习案例二：Iris clustering
    - （三）应用案例练习
- **三、超参数和模型验证**
    - （一）模型验证的错误方式
    - （二）模型验证的正确方式
        - 留出法（Holdout sets）
        - 交叉验证（Cross-validation）
    - （三）最佳模型的选择
        - 偏差（bias）-方差（variance）权衡
        - 验证曲线（validation curve）
        - 学习曲线（learning curves）
    - （四）实践中的验证：网格搜索（grid search）
- **四、特征工程（Feature engineering）**
    - （一）类别数据特征（categorical features）
    - （二）文本数据特征（text）
    - （三）图像数据特征 （image）
    - （四）衍生特征（derived features）
    - （五）缺失数据处理
- **五、朴素贝叶斯分类（Naive Bayes Classification）**
    - （一）贝叶斯分类算法
    - （二）高斯朴素贝叶斯（Gaussian Naive Bayes）
    - （三）多项式朴素贝叶斯（Multinomial Naive Bayes）

        - 案例：文本分类
    - （四）什么时候选择使用朴素贝叶斯？

Day 6 作业

https://www.datacamp.com/community/tutorials/the-importance-of-preprocessing-in-data-science-and-the-machine-learning-pipeline-i-centering-scaling-and-k-nearest-neighbours
阅读材料：
1. The "Python Machine Learning" book code repository and info resource https://github.com/rasbt/python-machine-learning-book
2. An Introduction to Statistical Learning (James, Witten, Hastie, Tibshirani, 2013) : Python code https://github.com/JWarmenhoven/ISLR-python
3. Building Machine Learning Systems With Python https://github.com/luispedro/BuildingMachineLearningSystemsWithPython

Day 7. 机器学习算法

- **一、线性回归**
    - （一）简单线性回归
    - （二）基函数回归（Basis Function Regression）
        - 多项式基函数（polynomial basis functions）
        - 高斯基函数（Gaussian basis functions）
    - （三）正则化（Regularization）
        - 岭回归（ridge regression）
        - 套索回归（lasso regression）
- **二、支持向量机（Support Vector Machines）**
    - （一）支持向量机的出现
        - 生成分类（Generative classification）
        - 判别分类（Discriminative classification）
    - （二）间隔最大化（Maximizing the Margin）：解决线性可分的分类问题
    - （三）核函数支持向量机（Kernel SVM）：解决线性不可分的分类问题
    - （四）优化支持向量机：Softening Margins
        - 案例：脸部识别技术
    - （五）小结
        - 支持向量机的优势与劣势
- **三、决策树和随机森林**
    - （一）决策树
        - 信息熵（Entropy）
        - 信息增益（Information gains）
        - 建立决策树
        - 过拟合（over-fitting）
    - （二）随机森林：估计量集成
        - 随机森林回归
    - （三）小结
        - 优势：快速训练和预测且可以同时进行；多个决策树使得概率分类成为可能；非参数模型非常灵活，可以很好地执行其他估计器拟合不足的任务。

- 劣势：结果很难被解释。

Day8：神经网络

- 神经网络简介：
  - 自然的神经网络→人造的神经网络：input layer, hidden layer, output layer。feedforward 与 backpropagation。用 A Neural Network Playground (tensorflow.org)进行直观的展示。
  - 了解构神经网络的 Batch, Iteration and Epoch 的关系以及实践中的区别。
  - 学习梯度下降算法，并自己动手操作用它回归一次；之后学习使用 torch 写好的梯度下降算法
  - Pytorch 的架构：通过拟合三个数据点的过程学习损失函数以及其如何反馈调整斜率与截距达到更好的拟合效果。
  - 学习激活函数 sigmoid 与 relu, 以及为什么要使用 Softmax 将输出映射到[0,1]，并确保和为 1。展示如何使用 pytorch 搭建一个多个隐藏层的神经网络（为什么要使用线性回归+逻辑斯蒂的方式？）
  - 实践操作：糖尿病患者分类以及 CIFAR10 数据集
- 手写数字识别
  - 使用正常的全链接神经网络效果差→使用 cnn 进行训练识别
- CNN
  - 卷积神经网络又称 CNN 是一种特殊的神经网络，主要用于处理网格状的数据形式（如时序数据为一维网格，图片数据为二维网络）。
  - 学习 CNN 的基本原理，以及实际操作中所需的流程。
  - CNN 的三个重要的理念：
    - Sparse interactions
    - Parameter sharing
    - Equivariant representations
  - 一个 CNN 的典型的三个阶段，以及池化的原理。
  - 案例：使用 pytorch 建立 CNN 并处理 MNIST 数据
- RNN
  - 循环神经网络又称 RNN 是一种用来处理序列数据的神经网络。（大多数循环神经网络也可以处理长度可变的序列）例如：时间序列、文字、音乐数据
  - RNN 简介：Teach RNN 'hihell' to 'ihello'
  - RNN with Embeddings（与 one-hot 编码方式进行对比）
  - 图解 RNN, LSTMs, GRU
- 利用 pytorch-forecasting 对时序数据进行预测
  - 这个包主要基于 TFT，一个可解释的应对多水平时间序列预测的模型。
  - 数据：Stallion & Co.'s global beer sales. 其有几千种独特的经销商与商品组合。使用此数据集训练并应用于新数据看其泛化能力。

Day8 作业：影评的情感分析

https://www.kaggle.com/c/sentiment-analysis-on-movie-reviews/data

参考资料

1. 神经网络：Code: https://github.com/hunkim/PyTorchZeroToAll
   Slides: http://bit.ly/PyTorchZeroAll
   Videos: https://www.bilibili.com/video/av15823922/

2. 深度学习：https://space.bilibili.com/88461692/channel/detail?cid=26587
https://pytorch.org/tutorials/beginner/pytorch_with_examples.html#pytorch-tensors
3. summer-school/class_03_卷积的直观理解.pdf at master · SocratesAcademy/summer-school (github.com)
4. https://poloclub.github.io/cnn-explainer/
5. https://karpathy.github.io/2015/05/21/rnn-effectiveness/
6. https://towardsdatascience.com/animated-rnn-lstm-and-gru-ef124d06cf45

参考文献

1. A. W. Harley, "An Interactive Node-Link Visualization of Convolutional Neural Networks," in ISVC, pages 867-877, 2015
2. B.Lim,S.Ö.Arık,N.Loeffetal.,Temporal Fusion Transformers for interpretable multi-horizon time series forecasting. International Journal of Forecasting (2021) https://doi.org/10.1016/j.ijforeca
3. *Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT press.*

Day9：word2vec

- 词嵌入引入
  - 文化的几何学
  - 词嵌入法可以在高维矩阵中将情感关系表示为几何关系。在这些向量空间中，代表词汇差异的矩阵的维度与文化意义的维度相一致（如男人-女人等）
  - HistWords

是一个工具与数据的集合，使用词嵌入法对语义变化进行分析。

  - 使用词嵌入量化 100 年来的性别与种族刻板印象

- Word2vec
  - Lead in:人格嵌入（什么样的人是相似的？）使用余弦相似计算两个有不同大五人格的人的相似性。
  - 基于谷歌新闻语料库训练的 word2vec，使用 Gensim 库获得该预训练模型，找到与某个词汇最相近的词，对不同的词矩阵进行加减运算。案例：对不同球类运动的阶级与性别偏向进行可视化。
  - 神经语言模型（如何根据给出的前两个单词预测下一个单词？马尔科夫）
  - 语言模型训练

连续词袋模型：基于预测目标前后的词进行预测
Skip-gram：基于当前的单词预测前后单词；如果只选择训练集中出现的单词效率低→需要从字典中随机抽取一些没有关系的词作为反例来训练。

  - Word2vec 的训练过程（CBOW 和 Skip-gram 两种形式）
  - 基于 pytorch 的 word2vec
  - 案例：NGram 词向量模型分析《三体》

参考资料

1. https://nlp.stanford.edu/projects/histwords/
2. https://jalammar.github.io/illustrated-word2vec/
3. http://mccormickml.com/2016/04/12/googles-pretrained-word2vec-model-in-python/
4. http://d2l.ai/chapter_natural-language-processing-pretraining/word2vec.html#the-skip-gram-model

参考文献

1. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
2. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural*

information processing systems, *26*.
3. Kozlowski, A. C., Taddy, M., & Evans, J. A. (2019). The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review*, *84*(5), 905-949.
4. Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. *arXiv preprint arXiv:1605.09096*.
5. Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, *115*(16), E3635-E3644.
6. Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183-186.
7. Bengio 2003 A Neural Probabilistic Language Model. Journal of Machine Learning Research. 3:1137–1155

Day10 自然语言处理初步&结课

- 自然语言处理
  - 文本处理步骤（标记词汇、特征向量）
    - 词袋模型的假设和优劣势（短语与词汇的关系、如何处理？）
    - 文档-术语矩阵 Document-Term Matrix (DTM)及其代码实现
    - 特征向量转换（n-gram 模型、TF-IDF 算法、L2-normalization）
    - 可视化
    - 主题模型及其代码实现
  - 情感分析（需要处理哪些类型的词？如何处理？）
    - 基于字典的情感分析
      - 分句情感词、程度词、否定词、感叹号
      - 情感词汇
    - 基于机器学习的情感分析（有哪些深度学习平台？）

参考文献：

1. Michel, J.-B., et al. (2011). Quantitative Analysis of Culture Using Millions of Digitized Books. Science, 331, 176–182.
2. Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. (2011). "Learning Word Vectors for Sentiment Analysis." The 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011).

推荐阅读：

1. movies reviews 情感分析 http://nbviewer.jupyter.org/github/rasbt/python-machine-learning-book/blob/master/code/ch08/ch08.ipynb
2. Sentiment analysis with machine learning in R
http://chengjun.github.io/en/2014/04/sentiment-analysis-with-machine-learning-in-R/
3. 使用 R 包 sentiment 进行情感分析
https://site.douban.com/146782/widget/notes/15462869/note/344846192/
4. 中文的手机评论的情感分析 https://github.com/computational-class/Review-Helpfulness-Prediction
5. 基于词典的中文情感倾向分析
https://site.douban.com/146782/widget/notes/15462869/note/355625387/

课程目标

1. 使用 Python 处理大规模数据
2. 可以理解和使用机器学习
3. 理解计算社会科学的基本逻辑、熟悉计算社会科学研究的经典研究
4. 掌握自然语言和网络科学的基本方法