要做的内容

1. 文本处理(基础数据提取、重命名工具、去掉代码注释(//和/\* \*/两种，要求长数据能力))
2. ★文本转换(2016.07.12 -> 12 Jul 2016)
3. 代数、矩阵计算(看起来最牛逼)
4. 数据处理(GEP改进)
5. 加入数据处理常用的方法，比如地理位置标注(也是个简单的PS)？
6. 发掘新的牛逼的应用

★ 参考trifacta，将基础功能转换成数据清洗模式

1. 鼠标选择分隔符，可以实现
2. 自动拆分成多列
3. 实现替换(手动输入即可)
4. Select \* !?contains [选中内容]
5. 显示每一列的最上方显示此列类别(来自数据或自己定义)、数量并从高到低排列
6. 可以选择正、逆序排列，
7. 如果是数字，可以选择均值、中位数、标准差、四分位数，包括几位小数、科学记数法等,可以挑出大于等于或小于均值的数据…
8. 选择凑满前n%的数据，
9. 可单选、多选选出其中几种标签(然后附加操作为仅保留、删除、局部排序(加入顺序)等)
10. 若选中缺失行(数量为0)，则可选择插入某个值或删除
11. 查看某一列的详情，可以绘制相应图表(对地理位置做特殊处理，还可选择显示在地图上)
12. 点击某列，可以选择继续应用PBE系统(将已有的string processor加进来)
13. …把所有常用方法都加进来

综述不应是材料的罗列，而是对亲自阅读和收集的材料，加以归纳、总结，做出评论和估价。并由提供的文献资料引出重要结论。一篇好的综述，应当是既有观点，又有事实，有骨又有肉的好文章。

**参考深度学习3大牛综述**

DL综述文章结构：

Abstract

前言(背景、成功应用案例、一句话概述进展情况)

监督学习(什么是监督学习，用到的思想、具体实现时的方法等)

反向传播来训练多层神经网络

卷积神经网络

使用深度卷积神经网络进行图像理解

分布式特征表示与语言处理

递归神经网络

深度学习的未来展望

Program synthesis发展进程及应用场景综述

Abstract

前言(背景描述、应用案例、小展望)

可参考**FlashMeta**的introduction，将program synthesis分为**Deductive Synthesis**，**Deductive Synthesis**，**Domain-specific Inductive Synthesis**这3个阶段。

(早期原型 2000年左右)

(近似产品：多项式拟合，GEP)

(近几年)发展的起点 基于规则(submit)

1. Substring
2. 数字处理
3. Colorful
4. 矩阵运算
5. 教辅程序

现状分析 基于概率模型，有一部分能应用深度学习

1. 2016机器学习NLP模型
2. ICLR2017 PS字符串预测
3. NPI+R-NPI 可理解为PS

趋势预测

(某企业声称效率达谷歌的100倍)

总结

参考文献